A peer-reviewed open-access journal
NeoBiota
Advancing research on alien species and biological invasions

NeoBiota 110: 1–20 (2026)
DOI: 10.3897/neobiota.110.198728

Software Description

A standardized workflow for compiling and harmonizing first records of non-native species

Tanara Renard Truong[1,2], Franz Essl[3], Hanno Seebens[1]

1 *Institute of Bioinformatics, Justus-Liebig-University Giessen, Giessen, Germany*
2 *Senckenberg Society for Nature Research, Senckenberg Biodiversity and Climate Research Centre, Frankfurt am Main, Germany*
3 *Division of BioInvasions, Macroecology and Global Change, University of Vienna, Vienna, Austria*

Corresponding author: Hanno Seebens (hanno.seebens@uni-giessen.de)

Abstract

Biological invasions are a major driver of biodiversity loss, and tracking non-native species introductions is essential for biodiversity monitoring and policy assessment. First records, defined as the earliest documented evidence of a species occurring in the wild outside its native range, provide a standardized metric to quantify introduction dynamics across regions and taxonomic groups. However, compiling and harmonizing such records from diverse sources remain challenging due to non-standardized terminology and heterogeneous reporting standards. Here, we describe a reproducible workflow for standardizing first records of non-native species from diverse sources. The workflow is implemented in R and is designed to ensure consistency in taxonomy, harmonize geographic locations, convert heterogeneous temporal expressions into four-digit years using explicit transformation rules, and standardize terms following Darwin Core terminology. A key innovation of the workflow is the assignment of a confidence level to each standardized first record, reflecting the uncertainty of the original temporal information. The workflow is open-source, modular, and customizable, allowing users to apply it to their own data or modify standardization rules. It is applied to produce v4.0 of the FirstRecords dataset, which contains nearly 100,000 first records of more than 30,000 taxa and is structured into linked tables covering actual records, taxonomy, locations, and data sources. This represents a 17.8% increase in the amount of available first records compared to the previous version, of which approximately 8% is due to the refinement of the workflow. By providing a transparent and automated approach to data harmonization, this workflow facilitates large-scale analyses of invasion patterns and indicator development, while maintaining data integrity.

Key words: Alien species, biological invasions, dataset, FAIR, invasive alien species, long-term trends, robust trends, software, standardization

OPEN ACCESS

Academic editor: Petr Pyšek
Received: 8 May 2026
Accepted: 24 September 2026
Published: 21 October 2026

Citation: Renard Truong T, Essl F, Seebens H (2026) A standardized workflow for compiling and harmonizing first records of non-native species. NeoBiota 110: 1–20. https://doi.org/10.3897/neobiota.110.198728

Copyright: © Tanara Renard Truong et al. This is an open access article distributed under terms of the Creative Commons Attribution License (Attribution 4.0 International – CC BY 4.0).

Introduction

Biological invasions are a major driver of biodiversity loss, contributing to ecosystem changes, species extinctions, and reduction in people's good quality of life worldwide (IPBES 2019, 2023a). The Kunming-Montreal Global Biodiversity Framework (GBF) explicitly recognizes invasive non-native species (also termed "invasive alien species") as a priority issue and calls for improved monitoring to support the achievement of its targets (CBD 2022). First records of species introductions, defined as the year of first documented evidence of a species occurring

in the wild outside its native range, allow to track introduction dynamics and inform such monitoring frameworks (Latombe et al. 2017; McGeoch et al. 2023b). First records have enabled key scientific insights, such as revealing that the global rate of first records of non-native species has continued to increase over the last two centuries without signs of saturation (Seebens et al. 2017), shedding light on processes of historical invasion dynamics (Bonnamour et al. 2021) and enabling predictions of future trends for major taxonomic groups and continents (Seebens et al. 2021). First records thus provide the foundation for conducting analyses on long-term trends, informing policy, and predicting future dynamics of biological invasions. However, these records are often reported in non-standardized formats, with varying taxonomic standards, temporal information, and geographic resolution. Manual harmonization is time-consuming, non-transparent, and error-prone (Chapman 2005; Maletic and Marcus 2005; Grenié et al. 2023). The generation of a centralized global source of first records would therefore benefit from the development of a software running data processing, harmonization, and integration automatically.

A global dataset of first records of non-native species was first compiled by Seebens et al. (2017). The original intention for the generation of this called FirstRecords dataset was to provide a basis for a comparative analysis of long-term trends of non-native species introductions. Indeed, this publication provided a first comprehensive comparison of trends of non-native species accumulation across taxonomic groups and continents, and highlighted the continuous increase in numbers of non-native species as well as the rise of rates of introduction worldwide (Seebens et al. 2017). Since its first release the FirstRecords dataset has been used in various subsequent analyses and reports, which raised attention also outside the academic community. For example, the dataset has been considered of interest to the public by the data provider "Our world in data" and integrated into their dashboards (Ritchie et al. 2022). Further, the dataset has been used for presenting the status of biological invasions in the Thematic Assessment Report on Invasive Alien Species and Their Control and the Global Assessment Report on Biodiversity and Ecosystem Services by the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) (IPBES 2019, 2023a) and considered as a "critical source of information" in the Invasive Alien Species Toolkit released by the Convention of Biological Diversity (CBD) (Convention on Biological Diversity and International Union for Conservation of Nature [IUCN] 2024). Thus, the dataset has been taken up and relied upon beyond its original intention, which further highlights the need for an open and transparent workflow for handling the underlying data to provide transparency and generate trust.

The first version of the dataset has been compiled in a collaborative effort of > 40 researchers. Various datasets have been gathered by the team providing information of first records of non-native species somewhere in the world. The type, format, and content of datasets were diverse coming from scientific publications, reports, online databases, and unpublished sources. These datasets often lack standards in the use of terminology, the definition of first records and non-native species, geographic regions, and taxonomic information, which required developing a workflow to process the individual datasets and records, to translate terms into a common vocabulary, and to integrate harmonized datasets into a single one. This necessitates making numerous decisions about treating information at various points of the process. Because conventional scientific articles rarely provide

space to document every data-processing decision in detail, the generation of the FirstRecords was not easily reproducible by others and therefore lacked transparency. However, transparency and reproducibility is essential for any analysis and particularly for informing policy to generate trust in the outcomes (Pooter et al. 2017; Franz and Sterner 2018). Achieving this requires that biodiversity data processing follows common standards in biodiversity research such as Darwin Core terminology (Wieczorek et al. 2012; Groom et al. 2019), which allows data to be documented, shared, and validated in a consistent and interoperable manner.

Here, we address this shortcoming by presenting the fully revised workflow to standardize first record data and to generate the FirstRecords dataset. The new version aligns with common standard biodiversity terminology and includes a consistent set of transformation rules for taxonomy, first records, and locations, which makes it easy to adjust the transformation rules where needed. Further, each first record is accompanied by a standardized confidence statement based on the precision of the original information. This enables users to apply different levels of temporal stringency when analysing introduction patterns. We further demonstrate the usability of the workflow by applying it to generate an extended and revised version of the FirstRecords dataset. The dataset is designed to complement other openly available resources on non-native species distributions and impacts, such as SInAS (Gómez-Suárez et al. 2025) or GIDIAS (Bacher et al. 2025), and provides a harmonized global baseline suitable for biodiversity monitoring and assessment.

Methods

The FirstRecords workflow standardizes and integrates first record data through a sequence of six distinct steps (described in detail below), each addressing a specific dimension of harmonization: dataset preparation, taxonomic standardization, temporal standardization, spatial standardization, terminology harmonization, and final dataset generation (Fig. 1). Each step is implemented as an independent R script that takes the output of the previous step as input, allowing individual components to be inspected, adapted, or rerun separately.

The workflow was implemented with R (v 4.5.1) (R Core Team 2025), using the packages data.table (v 1.17.8), rgbif (v 2.8.2), and tidyverse (v 2.0.0) (see Code availability).

As input, the workflow accepts a single non-standardised dataset (i.e., csv, xlsx, or ods file), wherein field names and terminology have not yet been mapped to a common data standard (e.g., Darwin Core). Data extraction from unstructured sources (e.g., free text, PDF documents, websites) is not supported; records must therefore be extracted manually from the original sources and structured in tabular form prior to submission.

At minimum, the input dataset must contain the following fields: (1) taxon name, (2) year of introduction or naturalization, and (3) locations. Records lacking any of these fields cannot be processed, and are therefore removed by the workflow. In addition, the workflow can handle information about (4) bibliographic citation and (5) access rights of each source, (6) ecological characteristics (habitat), (7) invasion status (occurrence, means and degree of establishment), and (8) pathway of introduction. Such information is processed if provided. Each record (row) must contain only one event of a taxon-year-location combination. Having multiple taxa or multiple locations in one row will cause the removal of this record. The

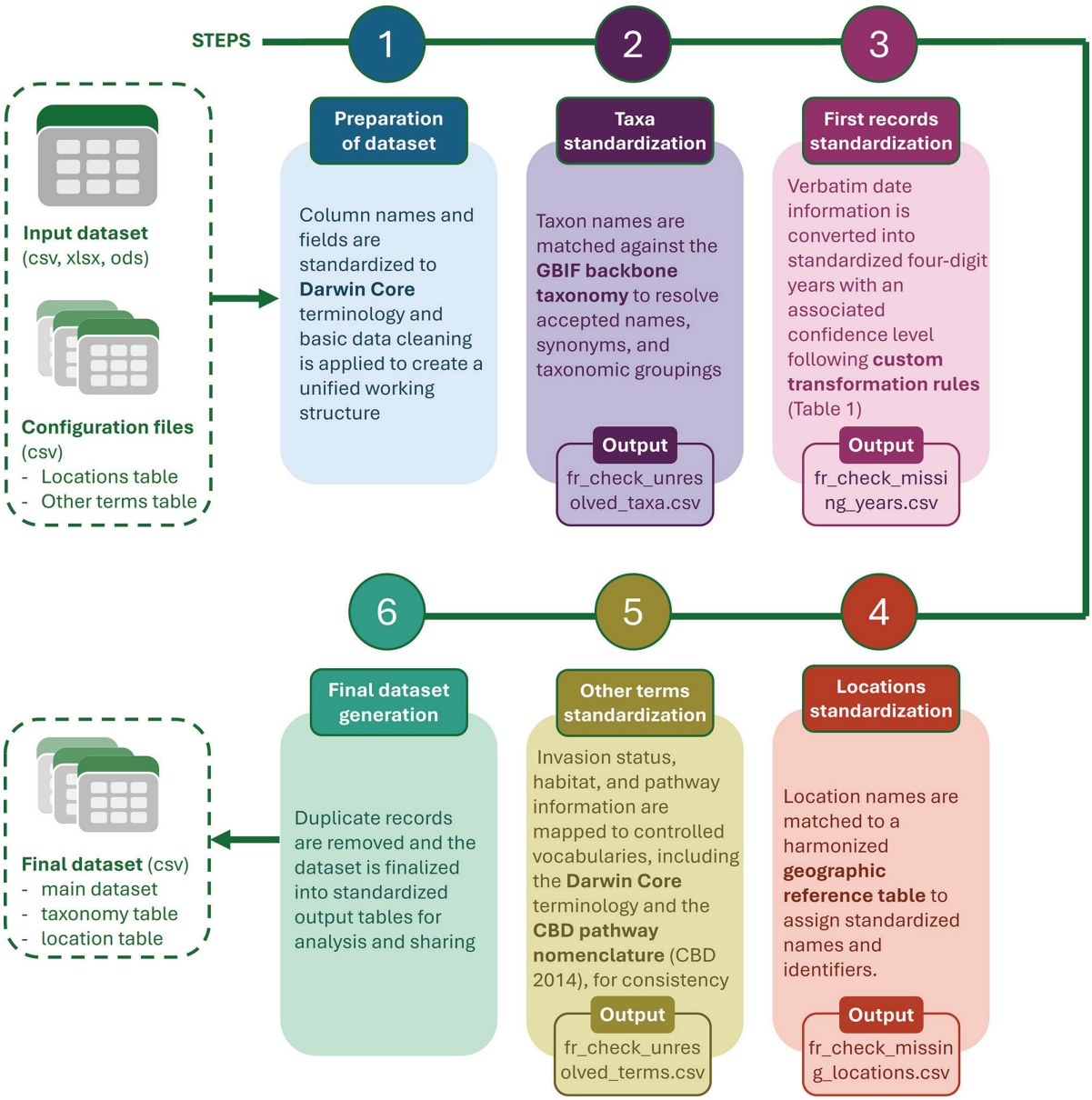

Figure 1. Overview of the standardized workflow for harmonizing first records of non-native species. The workflow comprises six sequential steps, each with defined inputs (gray boxes), processing procedures, and outputs (white boxes). Outputs from steps 2–5 are stored in a temporary folder for manual review, enabling the reference tables to be iteratively expanded based on user feedback. The workflow was implemented in R (v 4.5.1) and is publicly available on GitHub (see Code availability).

entry of the first record can have multiple years if, for example, ranges of years are provided, which should still refer to only one event.

The following subsections describe each step in detail, including the transformation rules applied and the outputs generated.

Step 1: Preparation of dataset

The initial phase of data processing involves the conversion of the column names of the input dataset to comply with the standard terminology for biodiversity data "Darwin Core" (Wieczorek et al. 2012; Groom et al. 2019), ensuring interoperability. Where possible, established Darwin Core terms are adopted directly (e.g., scientificName). However, because Darwin Core does not provide standard terms for

certain concepts specific to this workflow, custom fields were therefore introduced, following Darwin Core naming conventions where feasible. These include **firstRecordEvent**, **verbatimFirstRecordEvent**, and **confidenceFirstRecordEvent** as well as **taxon**, **location**, and **verbatimLocation**, which are simplified or renamed adaptations of related Darwin Core concepts (e.g., locality or verbatimLocality). Additional custom fields (e.g., **region**, **continent**, **taxaGroup**) were created to support the specific aggregation and taxonomic classification needs of this dataset. A basic data cleaning procedure is applied to all records, including the removal of unnecessary white spaces and the correction of minor formatting inconsistencies.

As part of the column harmonization and initial cleaning process, a temporary dataset structure is defined and initialized with relevant input fields, to facilitate data processing and integration. This intermediate structure builds on all relevant input fields required for subsequent standardization steps (see final structure described in Data Record). Notably, a **verbatimLocation** field is generated to store the originally recorded places of introduction, with a corresponding **location** field initialized with these values. Similarly, a **verbatimFirstRecordEvent** field is generated to store the earliest available date for each record, following an order of precedence: first occurrence in the wild is retained preferentially, and naturalization date is used only as a fallback when no such record is available. The **firstRecordEvent** field is then initialized with **verbatimFirstRecordEvent**. The **originalNameUsage** field was created by combining columns containing genus, species and authorship information from the input dataset. Other fields, including **scientificNameAuthorship**, **habitat**, **taxaGroup**, **occurrenceStatus**, **establishmentMeans**, **degreeOfEstablishment**, **pathway**, **datasetName**, **bibliographicCitation**, and **accessRights**, are directly inherited from the input dataset without prior modification. Other fields, including geographic ones (**locationID**, **country**, and **region**) are initialized as empty or "NA" values at this stage.

This step ensures that all subsequent operations are performed on a standardized dataset, facilitating reliable data filtering, merging, analysis, and generation of the final dataset.

This step is implemented in the R script fr_prepare_main_dataset.r, which contains all details of this step.

Step 2: Standardizing taxa

In the second step of the workflow, taxon names are standardized to ensure consistency of records. Input taxon names, including synonyms and unresolved taxon names are first cleaned by removing extra spaces and empty entries. Each taxon name is then checked against the GBIF backbone taxonomy (GBIF 2023) using the rgbif package to identify accepted names, synonyms, and unmatched taxa, and to correct minor spelling errors. The identification of accepted names is done through a series of checks, internally called "cases", implemented as a decision tree. For example, case 1 is defined as the exact match of the provided taxon name and the one provided in the GBIF backbone taxonomy. Case 2 refers to synonyms with exact matches, case 3 to fuzzy matches with high confidence, etc. The respective cases are exported in the final output table (**workflowNote**). This information helps to evaluate the quality of the match and to link the respective match to the decision tree in the workflow. Standardized taxon names were assigned to main taxon groups (**taxaGroup**; e.g., Mammals, Birds, Insects, Bryophytes, Algae,

Fungi, etc.; see Data overview) for easy access and aggregation. Unique identifiers (**taxonID**) were generated for each taxon (standardized and non-standardized), allowing consistent reference across the datasets. All records are kept, including those which could not be standardized at this step. The outputs of this step include the standardized main dataset including all records, a comprehensive, taxonomy table with GBIF status and taxon group assignments, and a list of unmatched taxa for further verification. This dataset is used as input for step 3.

This step is implemented in the R script fr_taxa_standard.r.

Step 3: Standardizing first records (years)

In the third step of the workflow, temporal information associated with first records (**firstRecordEvent**) is standardized to ensure consistent and comparable year data across sources. The process begins with text cleaning, including the removal of punctuation, special characters, and excess whitespace, and the conversion of number words into digits. Transformation rules used to standardize years are inspired by those developed by Dyer et al. (2017) and are summarized in Table 1. The following rules to standardize first records are applied (see Table 1 for examples): full dates are reduced to years, and any accompanying months or references are removed. Two-digit years are expanded to four digits, under the assumption that they refer to the 1900s. BC years are converted to negative values. Year ranges and decades are randomized within plausible intervals to avoid artificial peaks at e.g. the beginning of a decade. Centuries are translated into midpoint values. Other century expressions like early or late 19th century are converted to fixed years of the time interval (e.g., "early 19th century" to 1820). Relative temporal expressions are converted into numeric values relative to a reference year. The reference year is either defined as the year of the record publication for temporal expressions smaller than 100 years (e.g., "20 years ago"), or as the year 2000 for very uncertain temporal expressions (e.g., "150 years ago"). As a result, the standardized dataset retains only four-digit numeric year values. The field **verbatimFirstRecordEvent** in the final output allows the user to refer to the original first record (see Suppl. material 1: table S1).

Each standardized record is assigned a confidence level, reflecting the precision of the reported first record, rather than the accuracy of the underlying record itself. The assessment is based on the specificity of the temporal information, the width of temporal ranges, and historical plausibility. For example, high confidence is assigned if a single year after 1500 was provided in the original reference, while the record is of low confidence if it was taken from a wide year range. Table 1 provides a detailed summary of input formats, transformation rules, output examples, and confidence assignments applied to different types of temporal information. The outputs of this step include a cleaned dataset containing standardized years and corresponding confidence levels, as well as a list of unmatched or unprocessable entries for further verification. This dataset is used as input for step 4.

This step is implemented in the R script fr_years_standard.r.

Step 4: Standardizing locations

In the fourth step of the workflow, location information is standardized to ensure spatial consistency across records originating from different scales and sources. The workflow assumes one location per record; input records listing multiple locations

Table 1. Transformation rules for standardizing first records and assigning confidence terms.

Input format	Transformation rule	Input example	Output example	Confidence assignment
Exact year	Retained as numeric year	1991	1991	> 1500: high confidence ≤ 1500: low confidence
Dates	Extract year only	17 October 1991 October 1991	1991 1991	> 1500: high confidence ≤ 1500: low confidence
References attached	Strip reference, retain year	1817 (Smith, 1900)	1817	> 1500: high confidence ≤ 1500: low confidence
2-digit year	Expand to full year (default 1900s)	78	1978	> 1500: high confidence ≤ 1500: low confidence
Year ranges: two years separated by a hyphen, special character, word or space	Random number within the range	1817–1822 1817–22 1817/1822 1817 1822	1820 1818 1821 1818	≤1500: low confidence > 1500: confidence depends on input range width: ≤ 9 years: medium-high; 10–15 years: medium; 16–20 years: medium-low; > 20 years: low
Two years: two years separated by "and" or "&"	Assign earliest	1917 and 1920	1917	> 1500: high confidence ≤ 1500: low confidence
Two possible years: two years separated by "or"	Assign most recent	1917 or 1920	1920	> 1500: high confidence ≤ 1500: confidence depends on input range width: ≤ 9 years: medium-high; 10–15 years: medium; 16–20 years: medium-low; > 20 years: low
Decades	Randomized year within decade	1970s	1974	> 1500: medium-high confidence ≤ 1500: low confidence
Decade qualifiers (early, mid, late)	Convert to representative year (2, 5, or 8 of decade)	early 1970s mid 1970s late 1970s	1972 1975 1978	> 1500: medium-high confidence ≤ 1500: low confidence
Centuries	Convert to midpoint year	19th century	1850	low confidence
Century qualifiers	Convert to representative year (20, 50, 80 of century)	early 19th century mid 19th century late 19th century	1820 1850 1880	low confidence
Relative years (< 100 years ago)	Subtract from the year of publication	20 years ago (with source published in 2025)	2005	low confidence
Relative years (≥ 100 years ago)	Subtract from a reference year (set at 2000)	150 years ago	1850	low confidence
After/post years	Add 2 years to reference	after 1917	1919	> 1500: medium confidence ≤ 1500: low confidence
BC years	Converted to negative year	500 BC	-500	low confidence

(e.g., "France, Germany") are not split and are processed as a single compound string, which may affect downstream location standardization. Input data may include location names provided at various geographic levels (e.g., islands, regions, and countries) and in multiple languages or spellings. The standardization process

is borrowed from Gómez-Suárez et al. (2025), and relies on a reference table (AllLocations.xlsx; see Code availability) containing harmonized names, identifiers, and variant forms for global and subnational locations. Some subnational locations were retained based on their high relevance for invasion biology or geographic distance to the mainland (e.g., islands such as Hawaii). This reference table is distributed with the code. Geographic boundaries align with the Database of Global Administrative Areas dataset (GADM, version 2.8; https://gadm.org/), with national divisions corresponding to GADM's level 0 boundaries and sub-national divisions corresponding to level 1 or level 2 boundaries. Text fields are first cleaned by removing special characters, redundant whitespace, and language-specific articles (e.g., "the"), and all entries are converted to lowercase to facilitate matching. The script performs exact and keyword-based matching between the dataset and the reference tables of regions and subregions, accounting for synonyms, alternative spellings, and multilingual variants, as provided in the reference table. Each location is then assigned a standardized name and a unique **locationID**, linking it to a higher-level regional classification (e.g., IPBES sub-region). Locations that cannot be matched, including entities outside the scope of the reference table (e.g., seas, other non-administrative units or unrecognized names) are retained in the dataset but are not assigned a standardized location name or **locationID**; these records are also flagged and exported separately for manual review and potential inclusion in the reference table. Locations that match multiple regions or remain unresolved are flagged with warnings and listed for manual verification. The rationale for location assignment follows geographic relevance, prioritizing political or biogeographic units depending on the study context (e.g., island names, countries, administrative regions). The outputs of this step include a cleaned main dataset with standardized location identifiers and names, a comprehensive location reference table summarizing all recognized places and their associated regions, and a file containing unmatched or ambiguous entries for further curation.

The field **verbatimLocation** allows the user to refer to the original first record (see Suppl. material 1: table S1).

This step is implemented in the R script fr_localisation_standard.r.

Step 5: Standardizing remaining terms

In the fifth step of the workflow, invasion status-related attributes and pathway of introduction are standardized to comply with internationally recognized terminologies. In the input dataset, information related to invasion status (occurrence, means and degree of establishment) was collected under a single field, while habitats and pathways were recorded using different nomenclatures, spelling, or languages. The standardization process relies on several reference tables (e.g., reference_degreeOfEstablishment.csv, reference_habitat.csv, reference_establishmentMeans.csv, reference_occurrenceStatus.csv, reference_pathway.csv) containing harmonized terminology and variant forms for each term. The script performs exact and keyword-based matching between the dataset and the reference tables to each term category (occurrence status, establishment means, degree of establishment, and pathway), based on a controlled vocabulary of standardized terms and their associated verbatim variants (see configuration files in Code availability). For occurrence records, verbatim terms indicating non-presence (e.g., "extinct," "died out," "eradicated," "extirpated") are mapped to "absent"; terms indicating

confined or non-naturalized occurrence (e.g., "captive", "cultivated") are mapped to "present (not occurring in the wild)"; terms indicating ambiguous status (e.g., "not applicable," "unknown") are mapped to "uncertain."; and terms indicating presence (e.g., "present", "extant") are mapped to "present". Terms that do not match any entry in the reference tables are flagged and exported to a table file for manual review, allowing the reference table to be iteratively expanded. Following this matching step, any remaining empty entries are assumed to represent "present" occurrences, consistent with the tendency of sources to record occurrences explicitly only when a species is absent, not established in the wild, or of uncertain status. We recommend that users consult the unresolved-terms log to verify this assumption for their specific dataset before downstream analysis. Pathway standardization follows the nomenclature and hierarchical structure of the Convention on Biological Diversity (CBD 2014), including its categories and subcategories, using the same variant-matching and flagging approach. Habitat information is converted into "terrestrial", "freshwater", "brackish", and "marine" or combination of those, where available. Currently, habitat is the only ecological characteristic systematically standardized within the workflow, reflecting its prevalence and comparability across source datasets. Other ecological characteristics (e.g., trophic level or functional traits) are not currently included but could be incorporated in future versions by adding corresponding reference tables following the same matching approach used for existing terms. The resulting dataset provides a consistent set of categorical descriptors for each record, facilitating comparative analyses and data integration across scales.

Although not directly related to invasion status or ecological characteristics, the same approach is applied to streamline the **datasetName** field, ensuring uniform naming of sources of data sources across the workflow.

This step is implemented in the R script fr_terms_standard.r.

Step 6: Generating final dataset

In the final step, the processed dataset is cleaned, duplicates are removed, and exported as a standardized output file. Duplicate entries are first removed based on identical combinations of **taxon**, **verbatimLocation**, and **firstRecordEvent** to eliminate redundant records. The verbatim (rather than standardized) location field is used at this stage because location standardization (Step 4) only applies to a defined set of terrestrial administrative units matched against a reference table; locations that fall outside this scheme (e.g., seas, or other non-standardized entities) retain no standardized location value. The procedure then retains a single first record per taxon and standardized location, following a prioritization rule that favours entries derived from publications (non-empty **datasetName**), and when multiple such entries exist, selects the earliest record chronologically. The dataset is subsequently reordered by location and taxon for readability and consistency. The final FirstRecords dataset comprises four tables: a main data file, a taxonomy table, a location table, and a list of sources, available in .csv, and described in detail in the Suppl. material 1: tables S1–S5. The taxonomy table provides taxonomic information for each taxon, including its GBIF status and taxon group assignments. This step produces the definitive cleaned dataset used for downstream analyses and data sharing.

This step is implemented in the R script fr_save_output_dataset.r.

Application of the workflow

Input data compilation

Non-native species' first records are rarely documented in standalone publications, and therefore usually not mentioned in title, key words, or abstracts, which are the primary search criteria for systematic literature reviews. As a result, the collection of sources of first records necessitates a targeted, non-systematic approach, relying on the knowledge of the involved researchers. To address this challenge, the authors have aggregated first record data sources since 2015 with the support of their research network (see Acknowledgement), including now:

- 22 datasets or checklists of non-native species retrieved between 2013 and 2021, which cover various regions, scales (national to global) and taxonomic groups;
- 141 published peer-reviewed publications identified through continuous scientific monitoring (published between 1984 and 2025); and
- 22 personal communications obtained through professional networks and online data calls (received between 2013 and 2022).

Sources were reviewed, and relevant information compiled in a single excel spreadsheet manually, as first records are usually reported in highly variable, non-standardized formats (see Suppl. material 1: table S6). This ensured a rigorous quality control and consistent interpretation of metadata across heterogeneous sources. Only sources containing records of at least taxon name, year of introduction or naturalization, and locations were included in the input dataset.

The resulting input dataset consists of 221,771 records (taxon × region), including entries with restricted data-use permissions and before removal of duplicate or unusable records. It contains heterogeneous entries, with various taxonomic naming conventions (e.g., accepted names, synonyms, and sub-species), years (sometimes given as ranges, centuries, several years, or descriptive text), localities (from specific sites to continents), and other attributes (e.g., different pathway nomenclatures).

Standardization of first records and description of v4.0 of the internal FirstRecords dataset

The workflow was applied to the input file to produce a new version of the FirstRecords dataset (v4.0). After processing the input dataset by the workflow, the final internal FirstRecords dataset contains 98,020 standardized records of 30,279 taxa across 380 locations. It includes 8,622 records (8.8%) under restricted data permission, which we were not allowed to publish openly. Overall, 273 locations (71.8% of all locations in the original data sources) could be harmonized, aligning now with our location classification. Unstandardized locations (e.g., seas or multiple counties) are kept in the dataset, without any location identifier, and can also be found in the "location" table (see Data availability). Most represented continents in terms of number of records are Europe (39,155 records, 39.9%), Oceania (14,996, 15.3%) and North America (14,218, 14.5%) (Fig. 2). In total, 28,812 taxa (95.2% of all taxa) were harmonized, thus aligning with the GBIF backbone taxonomy. Most represented taxonomic groups are vascular plants (58,721, 59.9%), insects (17,637, 18%), and birds (4,279, 4.4%)

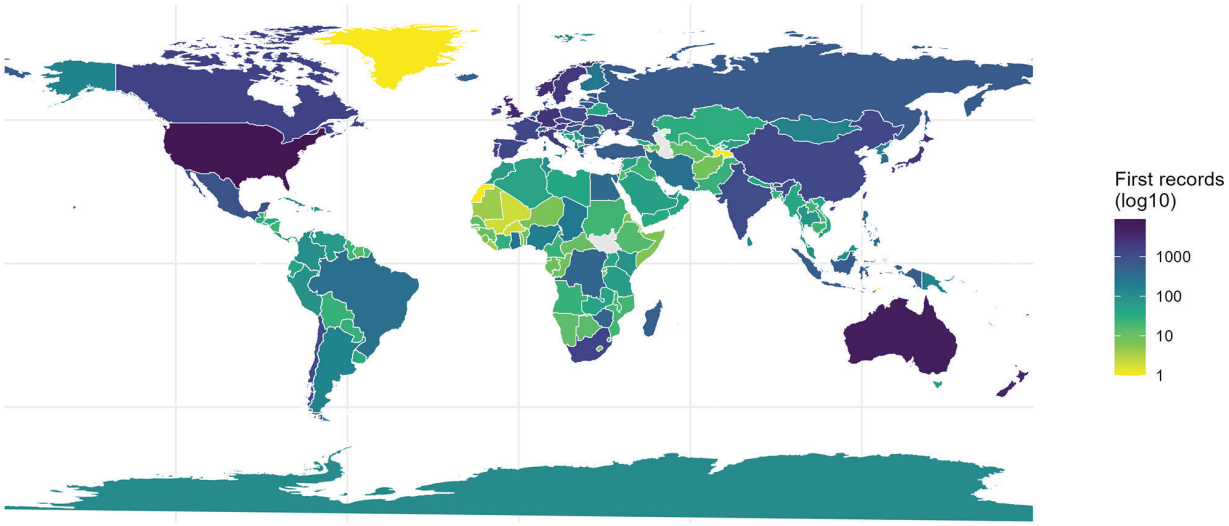

Figure 2. Global distribution of the total number of first records per location, shown on a log10 scale. Locations are defined in the location standardization step of the workflow (see Methods, Step 4). Colors correspond to the number of first records assigned to each location, with darker shades indicating higher counts. A logarithmic scale is used to accommodate the wide range in record counts across locations, from areas with very few documented first records to those with several hundred or more. Data underlying this figure are derived from the standardized output of the workflow described in this manuscript, including internal non-public records (see Data availability).

Table 2. Number of first records of non-native species per taxonomic group. 'Unresolved' corresponds to entries for which locations have not been standardized against the reference table (AllLocations.xlsx). SAR: Stramenopiles, Alveolates and Rhizaria.

Taxa\ Region	Africa	Asia	Europe	North America	Oceania	South America	Unresolved	Antarctica	Total
Algae	63	115	460	19	44	22	18	0	741
Amphibians	5	31	71	38	15	9	2	0	171
Annelids, nematodes, platyhelminthes, and other worms	55	148	769	122	52	24	69	21	1260
Arachnids	140	149	285	287	242	164	34	34	1335
Bacteria and protozoans	2	30	139	14	3	1	13	0	202
Birds	466	804	1247	798	737	110	103	14	4279
Bryophytes	3	1	69	22	31	31	35	2	194
Crustaceans	159	264	1092	151	145	81	129	2	2023
Fishes	462	1146	1168	538	189	162	146	15	3826
Fungi	32	39	502	60	147	112	5	19	916
Insects	1006	1100	5735	2748	3088	859	2942	173	17651
Mammals	80	62	638	109	122	101	24	80	1216
Molluscs	140	479	942	151	224	57	108	7	2108
Other	16	68	306	19	203	18	27	0	657
Other aquatic animals	117	120	552	104	156	80	85	0	1214
Other arthropods	8	1	103	4	8	17	17	36	194
Reptiles	25	45	152	90	23	14	2	0	351
SAR	18	137	614	25	26	11	29	0	860
Vascular plants	4948	7865	24224	8920	9542	1553	981	668	58701
Viruses	3	11	83	3	0	0	0	0	100

(Table 2). A total of 37,903 (38.7%) records present information on their degree of establishment, 98,020 (100%) on their occurrence status, 40,537 (41.4%) on their establishment mean, and 1,938 (2%) on their introduction pathway.

Furthermore, information on the habitat is available for 3,409 taxa (3.5%). Most first records are high-confidence (85,153, 86.9%), with a minor proportion of records with medium-high confidence (7,2–6, 7.4%) medium (379, 0.4%), medium-low (130, 0.1%) and low (5,152, 5.2%) confidence.

The number of first records per year increased substantially over the study period, revealing a long-term acceleration in the documentation of new occurrences (Fig. 3). From 1800 onwards, annual counts remained relatively low and stable, typically below 250 records per year, before entering a period of gradual increase through the late 19th and early 20th centuries, as was also described by Seebens et al. (2017). A marked acceleration is evident from approximately 1950, culminating in a peak of more than 1000 first records per year around 2000. This is followed by a sharp decline in the most recent years, reflecting a reporting lag rather than a genuine decrease in new occurrences, as recent records may not yet have been fully digitised, verified, or submitted to the database.

The close tracking between all confidence levels (orange) and high confidence records (teal) throughout most of the time series suggests that the overall trend is robust and not driven by poorly documented or uncertain records (Fig. 3). The proportion of non-high confidence records is decreasing over time, from 0.25 in 1800 to 0.13 in 2000 (Fig. 4), probably due to improved reporting standards in recent years.

Version 4.0 of the dataset represents a substantial expansion compared to version 3.1, with a 17.8% increase in the number of records, including notable increases in both taxonomic coverage (+15% unique taxa) and spatial coverage (+27% unique

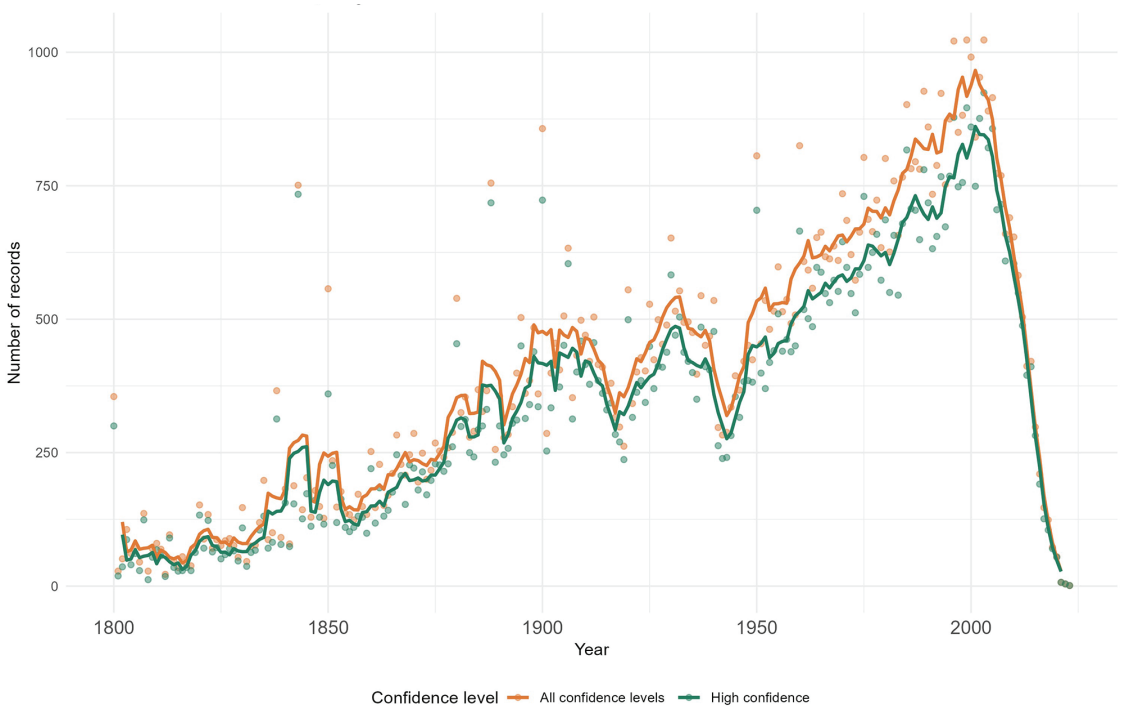

Figure 3. Time series, number of first records per year (1800–2023). Points represent the yearly counts of first records, while the solid line shows the 5-year moving average, smoothing short-term fluctuations to highlight longer-term trends. Note that counts in recent years are likely to be underestimated due to reporting lag (Seebens et al. 2017), between the time delay between a species' actual first occurrence and its documentation and publication in the sources compiled here. The apparent decline in first records in the most recent years should therefore not be interpreted as a genuine decrease in introduction rates, but rather as an artifact of incomplete or delayed reporting. Data underlying this figure are derived from the standardized output of the workflow described in this manuscript, including internal non-public records (see Data availability).

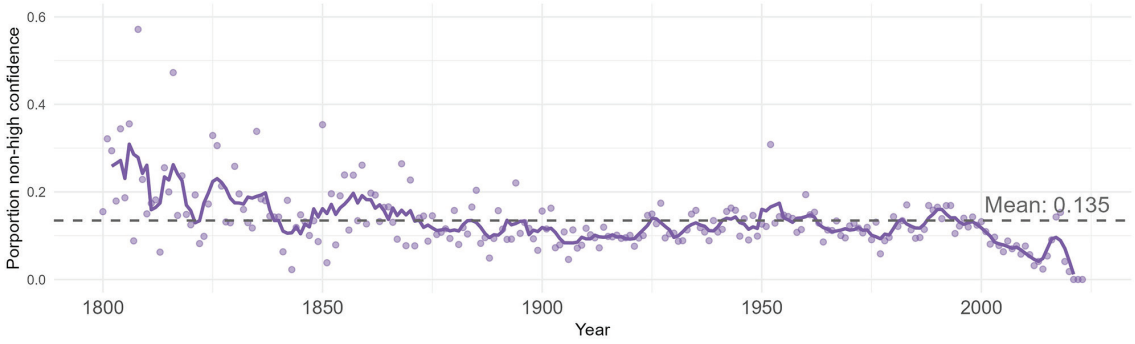

Figure 4. Temporal trend in the proportion of records assigned non-high confidence levels over time (1800–2023). Points represent the yearly proportion of non-high confidence records (i.e., medium-high, medium-low or low) among all first records reported that year, while the solid line shows the 5-year moving average, smoothing short-term fluctuations to highlight longer-term trends. The overall trend reflects changes in the specificity and reliability of the temporal information available across historical sources, rather than changes in the true timing of species introductions. Data underlying this figure are derived from the standardized output of the workflow described in this manuscript, including internal non-public records (see Data availability).

locations). The most dominant group in both versions is vascular plants, which grew by nearly 23%, further consolidating its dominance. Insects and molluscs also saw meaningful increases in record numbers (12% and 14% respectively), and bacteria and protozoans stood out with the largest proportional increase (42%), albeit from a small baseline. However, not all groups expanded: mammals, reptiles, and amphibians saw considerable declines in record numbers (around 27%, 31%, and 40% respectively), as did algae (39%) due to modifications of the harmonization process. Approximately 8% of additional records are due to the refinement of the workflow, but these changes may also reflect reclassification in taxonomic groups (e.g., the usage of stramenopiles, alveolates, and rhizarians (SAR) resulted in a decline of records for Algae) between versions as well as removal of uncertain records and additions of new records in the input dataset of version 4.0.

Discussion

The workflow presented here constitutes a substantial methodological progress compared to previous approaches to compiling and harmonizing first records of non-native species. It provides a fully documented, open-source, and modular automated pipeline for standardization. By explicitly encoding all processing decisions, from taxonomic matching rules to temporal transformation rules, the workflow transparently supports replication. Furthermore, alignment with Darwin Core terminology and other biodiversity standards (e.g., taxonomy) allows the workflow to produce interoperable datasets in line with major global biodiversity infrastructure, including GBIF, which also complies with Target 21 of the KM GBF. Finally, the modular architecture and automation of the workflow mean that individual steps can be updated or replaced independently as standards evolve or new reference databases become available, considerably reducing the time and technical effort required to update the dataset relative to the previous, compilation process – though expert curation remains essential for source compilation and for resolving ambiguous taxonomic or geographic entries.

The development of standardized, reproducible workflows is increasingly recognized as a critical need in invasion science, where heterogeneous data sources and inconsistent processing decisions can substantially affect analytical outcomes and

cross-study comparability (Griffith et al. 2026). Recent initiatives (Vanderhoeven et al. 2017; Saffer et al. 2024; Gómez-Suárez et al. 2025) illustrate this growing momentum: the DASCO workflow (Seebens and Kaplan 2022) provides a reproducible process for downscaling and converting non-native species checklists across different spatial units, while the SInAS workflow (Seebens et al. 2020) provides a standardized approach for assembling and harmonizing invasion data from multiple sources. Together, these efforts highlight how workflow-based approaches can reduce arbitrariness in data processing, improve transparency, and facilitate the integration of datasets across research groups and policy contexts. The workflow presented here builds on this trajectory by addressing the specific challenges associated with first record data, including temporal uncertainty and heterogeneous geographic naming conventions, contributing to an emerging ecosystem of interoperable tools for invasion science.

A key novelty of the workflow is the systematic assignment of confidence levels to standardized temporal information. Many sources report first records with approximate temporal information, such as "18th century" or "early 1970s", which reflect the uncertainty in historical documentation. The introduced confidence level system, ranging from high to low, makes this inherent uncertainty explicit and standardized, providing users with an option for filtering records according to their analytical requirements. For trend analyses, rate estimation, and the derivation of policy-relevant indicators at the local or regional scale, restricting analyses to records of high or medium-high confidence is advisable, as these records are based on specific or near-specific year information. Broader analysis, by contrast, may benefit from incorporating the full dataset, including low-confidence entries, provided that the potential for temporal uncertainty is acknowledged.

Despite the increase in the number of records, taxonomic and geographic imbalances persist in v4.0 of the FirstRecords dataset, with vascular plants remaining the most documented group, reflecting taxonomic biases in biodiversity data in general (McRae et al. 2017; Hughes et al. 2021; García-Roselló et al. 2023). Likewise, Europe and North America are being substantially better represented than Africa, South America, and parts of Asia across nearly all taxonomic groups, mirroring the global distribution of invasion science capacity and monitoring infrastructure (IPBES 2023a; Seebens et al. 2025). These biases should be interpreted carefully as it likely reflects research and reporting effort rather than true geographic differences in introduction rates.

Limitations to the workflow

The workflow achieves a high degree of automation and consistency across standardization steps, while several areas remain where user input and expert judgment continue to play an important role. Taxonomic matching relies on the GBIF Backbone Taxonomy which, while comprehensive, is not exhaustive and may not reflect the most current accepted names for all groups, particularly for poorly studied taxa. Fuzzy matching of taxon names reduces the number of unresolved entries but may occasionally introduce incorrect synonymization, particularly for older or informal names. The list of unmatched taxa produced at each workflow run therefore requires manual verification, which remains a resource-intensive step.

Location standardization is similarly subject to the scope of the predefined reference lists used for place-name matching. Unusual, historical, or sub-national location names may remain unresolved, and the prioritization rules governing the assignment

of records to political or biogeographic units may not be appropriate for all analytical contexts. In particular, records associated with island names, seas, and oceans may benefit from manual curation to complement what the automated workflow provides. Users applying the workflow to datasets with substantial geographic heterogeneity should therefore review the list of unmatched locations produced at step 4, and may supplement the reference location table with domain-specific entries as needed.

Further, although temporal information is standardized following explicit transformation rules (see Methods), some year values involve random assignment within predefined intervals to avoid artificial peaks. A random seed has been set in the workflow to ensure reproducibility, but users are invited to refer to verbatim first records (**verbatimFirstRecordEvent**) and assigned confidence level (**confidenceFirstRecordEvent**) for more specific information.

A related consideration concerns which parts of the workflow are automated and which still depend on human input. Of the six standardization steps, four (preparation of dataset, standardization of first records, terminology harmonization, and final dataset generation; steps 1, 3, 5, and 6) are fully automated, requiring no manual intervention once the R scripts are executed. The remaining two (taxonomic and location standardization; steps 2 and 4) are automated, allowing expert review of unmatched taxa or ambiguous location names. However, compiling first-record sources still requires expert knowledge, systematic literature monitoring, and manual data entry, none of which have been automated. Relative to the previous approach, the current workflow substantially reduces the time and effort required once source data are compiled, with a full re-run of the six standardization steps taking minutes rather than the hours previously required. However, it does not reduce the expertise or effort needed to locate and compile new first-record sources.

Limitations of the input dataset

Users should be aware of several important limitations when working with the FirstRecords dataset, many of which reflect broader challenges in biological invasion research. The temporal and spatial coverage exhibits considerable variation across regions, taxonomic groups, and source materials (IPBES 2023b; Seebens et al. 2025), potentially introducing sampling biases that may affect analyses (Seebens et al. 2017). First record dates represent documented introductions rather than actual arrival times, which may introduce a time lag (Essl et al. 2011). Where precise geographic information was unavailable, some records retain broader spatial resolutions (e.g., country-level). Additionally, the current workflow does not include a formal mechanism for assessing the reliability of individual sources or records. The confidence levels assigned to records reflect the precision of temporal information as reported in the original source; they do not capture uncertainty arising from the quality or credibility of the source itself. This distinction is important: a record reported with a precise four-digit year will receive a high confidence level regardless of the nature of the underlying evidence. Future versions of the workflow could explore the integration of source-level quality indicators with, for example, distinction between peer-reviewed publications from personal communications or unpublished reports, to provide users with a more complete picture of record reliability and to further strengthen the interpretive value of the confidence level system.

Finally, the full dataset includes entries with restricted usage permissions, flagged in the **accessRights** field, to ensure compliance with data-sharing agreements and

ethical considerations in biodiversity research (FAIR Data Principles). As we are not allowed to share datasets with restricted permissions, these are excluded from the openly available version on Zenodo. These limitations align with broader challenges in invasion science, where data heterogeneity, detection biases, and reporting inconsistencies are widely recognized (McGeoch et al. 2023a).

Conclusions

The standardized workflow presented here represents a significant methodological advancement in the compilation and harmonization of first records of non-native species. It complements other existing workflows in providing transparent, automated, and reproducible approaches to process data. Although often underrated, such workflows provide essential components for making science open and FAIR. Thus, the release and publication of workflows should become standard practice in scientific analyses also to enable other researchers to re-use components. Further, the publication of workflows generates trust in data processing and resulting datasets, which are then better positioned to contribute to the development of strategies and indicators relevant to international policy frameworks, including Target 6 of the Kunming-Montreal GBF. However, continued efforts to expand geographic and taxonomic coverage, reduce detection biases, and improve data quality remain essential to continue enhancing the dataset's accuracy and policy relevance.

Code availability

The standardization workflow is open-source and available on Github at https://github.com/EcoInf-group/FirstRecords, including:

- R scripts for all steps (R/).
- Configuration files (data/config/), which define standard locations (regions/sub-regions) and controlled vocabularies for **occurrenceStatus**, **establishmentMeans**, **degreeOfEstablishment**, **pathway**, and **habitat**. These files are modular and can be replaced or adapted to user-specific needs (e.g., alternative regional or taxonomic standards).
- Example input data (data/input/) corresponding to an extract of version v4.0 of the full FirstRecords dataset (see Data availability). Users may also replace this file with their own data, provided that it contains the same column structure. This sample dataset is also included as a supplementary material to this manuscript (Suppl. material 1: table S6).
- Users can execute the standardization workflow by opening the RStudio project file (FirstRecord.Rproj) and running the main script (R/fr_main_workflow.R).

Data availability

The dataset (version 4.0), generated using this workflow, is available on Zenodo (Seebens and Renard Truong 2026): https://doi.org/10.5281/zenodo.3690741.

The repository includes the main output "FirstRecords_dataset_public_4.0.csv" along with the taxonomy table "FirstRecords_taxonomy_table_public_4.0.csv" and the location table "FirstRecords_location_table_public_4.0.csv". It also

includes the input dataset "FirstRecords_raw_dataset_public_4.0.csv" and the list of sources "FirstRecords_dataset_sources_public_4.0.csv".

Note that this dataset does not include records with restricted usage permission. The list of sources is also included as supplementary material to this manuscript (Suppl. material 1).

Acknowledgements

We are grateful to all contributors for the provision of datasets of first records and for continuous support in identifying new datasets. The contributors are listed as "Data collectors" on the Zenodo webpage of the FirstRecords dataset (see Data availability). We are thankful for the feedback of two anonymous reviewers and the handling editor (Petr Pyšek).

References

Bacher S, Ryan-Colton E, Coiro M, Cassey P, Galil BS, Nuñez MA, Ansong M, Dehnen-Schmutz K, Fayvush G, Fernandez RD, Hiremath AJ, Ikegami M, Martinou AF, McDermott SM, Preda C, Vilà M, Weyl OLF, Aravind NA, Angelidou I, Athanasiou K, Atkore V, Barney JN, Blackburn TM, Brockerhoff EG, Carbutt C, Carisio L, Castro-Díez P, Céspedes V, Christopoulou A, Cisneros-Heredia DF, Cooling M, de Groot M, Demetriou J, Dickey JWE, Duboscq-Carra VG, Early R, Evans TE, Flores-Males PT, Gallardo B, Gruber M, Hui C, Jeschke JM, Joelson NZ, Khan MA, Kumschick S, Lach L, Lapin K, Lioy S, Liu C, MacMullen ZJ, Mazzitelli MA, Measey J, Mrugała-Koese AA, Musseau CL, Nahrung HF, Pepori A, Pertierra LR, Pienaar EF, Pyšek P, Rivas Torres G, Rojas Martinez HA, Rojas-Sandoval J, Ryan-Schofield NL, Sánchez RM, Santini A, Santoro D, Scalera R, Schmidt L, Shivambu TC, Sohrabi S, Tricarico E, Trillo A, van't Hof P, Volery L, Zengeya TA (2025) Global Impacts Dataset of Invasive Alien Species (GIDIAS). Scientific Data 12: 832. https://doi.org/10.1038/s41597-025-05184-5

Bonnamour A, Gippet JMW, Bertelsmeier C (2021) Insect and plant invasions follow two waves of globalisation. Ecology Letters 24(11): 2418–2426. https://doi.org/10.1111/ele.13863

CBD (2014) Pathways of introduction of invasive species, their prioritization and management. https://www.cbd.int/doc/meetings/sbstta/sbstta-18/official/sbstta-18-09-add1-en.pdf

CBD (2022) Decision 15/4: Kunming-Montreal Global Biodiversity Framework. https://www.cbd.int/gbf/targets/6

Chapman AD (2005) Principles and Methods of Data Cleaning: Primary Species and Species-Occurrence Data, version 1.0. Global Biodiversity Information Facility, Copenhagen. http://www.gbif.org/document/80528

Convention on Biological Diversity and International Union for Conservation of Nature [IUCN] (2024) Invasive Alien Species Toolkit for Target 6 of the Kunming-Montreal Global Biodiversity Framework. The Secretariat of the Convention on Biological Diversity.

Dyer EE, Redding DW, Blackburn TM (2017) The global avian invasions atlas, a database of alien bird distributions worldwide. Scientific Data 4: 170041. https://doi.org/10.1038/sdata.2017.41

Essl F, Dullinger S, Rabitsch W, Hulme PE, Hülber K, Jarošík V, Kleinbauer I, Krausmann F, Kühn I, Nentwig W, Vilà M, Genovesi P, Gherardi F, Desprez-Loustau M-L, Roques A, Pyšek P (2011) Socioeconomic legacy yields an invasion debt. Proceedings of the National Academy of Sciences 108: 203–207. https://doi.org/10.1073/pnas.1011728108

Franz NM, Sterner BW (2018) To increase trust, change the social design behind aggregated biodiversity data. Database 2018: bax100. https://doi.org/10.1093/database/bax100

García-Roselló E, González-Dacosta J, Lobo JM (2023) The biased distribution of existing information on biodiversity hinders its use in conservation, and we need an integrative approach to act urgently. Biological Conservation 283: 110118. https://doi.org/10.1016/j.biocon.2023.110118

GBIF (2023) GBIF Backbone Taxonomy. https://doi.org/10.15468/39omei

Gómez-Suárez M, Laeseke P, Seebens H (2025) A global dataset of native and alien distributions of alien species. Scientific Data 12: 1914. https://doi.org/10.1038/s41597-025-06379-6

Grenié M, Berti E, Carvajal-Quintero J, Dädlow GML, Sagouis A, Winter M (2023) Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. Methods in Ecology and Evolution 14: 12–25. https://doi.org/10.1111/2041-210X.13802

Griffith J, Lord J-M, Catchen MD, Arce-Plata MI, Blanchet FG, Chandramohan M, Diaz-Corzo MC, Gravel D, Gutiérrez C, Helfenstein IS, Hoban S, Kass JM, Laikre L, Larocque G, Leigh DM, Leung B, Mastretta-Yanes A, Millette KL, Berbeo MAM, Nguyen D, Norman KE, Olaya-Rodríguez MH, Pahls S, Pereira K, Peres-Neto PR, Poisot T, Pollock LJ, Rey-Velasco JC, Rincon-Parra VJ, Roeoesli C, Rousseu F, Sánchez-Clavijo LM, Schuman MC, Selmoni O, da Silva JM, Suarez-Valencia E, Surasinghe TD, Turak E, Urbina LF, Valentin S, Wightman N, Zuloaga J, Londoño MC, Gonzalez A (2026) BON in a Box: An Open and Collaborative Platform for Biodiversity Monitoring, Indicator Calculation, and Reporting. BioScience 76: 345–358. https://doi.org/10.1093/biosci/biaf189

Groom Q, Desmet P, Reyserhove L, Adriaens T, Oldoni D, Vanderhoeven S, Baskauf SJ, Chapman A, McGeoch M, Walls R, Wieczorek J, Wilson J, Zermoglio PF, Simpson A (2019) Improving Darwin Core for research and management of alien species. Biodiversity Information Science and Standards 3: e38084. https://doi.org/10.3897/biss.3.38084

Hughes AC, Orr MC, Ma K, Costello MJ, Waller J, Provoost P, Yang Q, Zhu C, Qiao H (2021) Sampling biases shape our view of the natural world. Ecography 44: 1259–1269. https://doi.org/10.1111/ecog.05926

IPBES (2019) Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Brondizio E, Díaz S, Settele J, Ngo HT (Eds). Zenodo. https://doi.org/10.5281/zenodo.6417333

IPBES (2023a) IPBES Invasive Alien Species Assessment: Full report. Roy HE, Pauchard A, Stoett P, Renard Truong T (Eds). https://doi.org/10.5281/zenodo.11629357

IPBES (2023b) IPBES Invasive Alien Species Assessment: Summary for Policymakers. Roy HE, Pauchard A, Stoett P, Renard Truong T, Bacher S, Galil BS, Hulme PE, Ikeda T, Sankaran K, McGeoch MA, Meyerson LA, Nuñez MA, Ordonez A, Rahlao SJ, Schwindt E, Seebens H, Sheppard AW, Vandvik V (Eds). https://doi.org/10.5281/zenodo.11254974

Latombe G, Pyšek P, Jeschke JM, Blackburn TM, Bacher S, Capinha C, Costello MJ, Fernández M, Gregory RD, Hobern D, Hui C, Jetz W, Kumschick S, McGrannachan C, Pergl J, Roy HE, Scalera R, Squires ZE, Wilson JRU, Winter M, Genovesi P, McGeoch MA (2017) A vision for global monitoring of biological invasions. Biological Conservation 213: 295–308. https://doi.org/10.1016/j.biocon.2016.06.013

Maletic JI, Marcus A (2005) Data Cleansing. In: Maimon O, Rokach L (Eds) Data Mining and Knowledge Discovery Handbook. Springer US, Boston, MA, 21–36. https://doi.org/10.1007/0-387-25465-X_2

McGeoch MA, Ordonez A, Howard PL, Groom QJ, Shrestha BB, Fernandez Trigoso M, Brugnoli E, Bwalya B, Byun C, Ksenofontov S, Ojaveer H, Simberloff D, Mungi NA, Rono B (2023a) Chapter 6. Governance and policy options for the management of biological invasions. In: Roy HE, Pauchard A, Stoett P, Renard Truong T (Eds) Thematic Assessment Report on Invasive Alien Species and their Control of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Zenodo. https://doi.org/10.5281/zenodo.10677227

McGeoch MA, Buba Y, Arlé E, Belmaker J, Clarke DA, Jetz W, Li R, Seebens H, Essl F, Groom Q, García-Berthou E, Lenzner B, Meyer C, Vicente JR, Wilson JRU, Winter M (2023b) Invasion

trends: An interpretable measure of change is needed to support policy targets. Conservation Letters 16: e12981. https://doi.org/10.1111/conl.12981

McRae L, Deinet S, Freeman R (2017) The diversity-weighted living planet index: Controlling for taxonomic bias in a global biodiversity indicator. PLOS ONE 12: e0169156. https://doi.org/10.1371/journal.pone.0169156

Pooter DD, Appeltans W, Bailly N, Bristol S, Deneudt K, Eliezer M, Fujioka E, Giorgetti A, Goldstein P, Lewis M, Lipizer M, Mackay K, Marin M, Moncoiffé G, Nikolopoulou S, Provoost P, Rauch S, Roubicek A, Torres C, van de Putte A, Vandepitte L, Vanhoorne B, Vinci M, Wambiji N, Watts D, Salas EK, Hernandez F (2017) Toward a new data standard for combined marine biological and environmental datasets – expanding OBIS beyond species occurrences. Biodiversity Data Journal 5: e10989. https://doi.org/10.3897/BDJ.5.e10989

R Core Team (2025) R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Ritchie H, Spooner F, Roser M (2022) Data Page: Global first records of non-native species – "Biodiversity". Data adapted from Seebens et al. Our World in Data. https://archive.ourworldindata.org/20260518-090244/grapher/global-first-records-of-established-alien-species.html [July 9, 2026]

Saffer A, Worm T, Takeuchi Y, Meentemeyer R (2024) GIATAR: a Spatio-temporal Dataset of Global Invasive and Alien Species and their Traits. Scientific Data 11: 991. https://doi.org/10.1038/s41597-024-03824-w

Seebens H, Kaplan E (2022) DASCO: A workflow to downscale alien species checklists using occurrence records and to re-allocate species distributions across realms. NeoBiota 74: 75–91.

Seebens H, Renard Truong T (2026) FirstRecords: A global dataset of first records of non-native species. https://doi.org/10.3897/neobiota.74.81082

Seebens H, Blackburn TM, Dyer EE, Genovesi P, Hulme PE, Jeschke JM, Pagad S, Pyšek P, Winter M, Arianoutsou M, Bacher S, Blasius B, Brundu G, Capinha C, Celesti-Grapow L, Dawson W, Dullinger S, Fuentes N, Jäger H, Kartesz J, Kenis M, Kreft H, Kühn I, Lenzner B, Liebhold A, Mosena A, Moser D, Nishino M, Pearman D, Pergl J, Rabitsch W, Rojas-Sandoval J, Roques A, Rorke S, Rossinelli S, Roy HE, Scalera R, Schindler S, Štajerová K, Tokarska-Guzik B, Van Kleunen M, Walker K, Weigelt P, Yamanaka T, Essl F (2017) No saturation in the accumulation of alien species worldwide. Nature Communications 8: 14435. https://doi.org/10.1038/ncomms14435

Seebens H, Clarke DA, Groom Q, Wilson JRU, García-Berthou E, Kühn I, Roigé M, Pagad S, Essl F, Vicente J, Winter M, McGeoch M (2020) A workflow for standardising and integrating alien species distribution data. NeoBiota 59: 39–59. https://doi.org/10.3897/neobiota.59.53578

Seebens H, Bacher S, Blackburn TM, Capinha C, Dawson W, Dullinger S, Genovesi P, Hulme PE, van Kleunen M, Kühn I, Jeschke JM, Lenzner B, Liebhold AM, Pattison Z, Pergl J, Pyšek P, Winter M, Essl F (2021) Projecting the continental accumulation of alien species through to 2050. Global Change Biology 27: 970–982. https://doi.org/10.1111/gcb.15333

Seebens H, Meyerson LA, Richardson DM, Lenzner B, Tricarico E, Courchamp F, Aleksanyan A, Keskin E, Saeedi H, Akite P, Alexander JM, Bailey SA, Biancolini D, Blackburn TM, Boehmer HJ, Bortolus A, Cadotte MW, Capinha C, Carlton JT, Crouch JA, Daehler CC, Essl F, Foxcroft LC, Fridley JD, Fuentes N, Gaertner M, Galil B, García-Berthou E, García-Díaz P, Haider S, Heneghan L, Hughes KA, Hui C, Kaplan E, Liebhold AM, Liu C, Marchante E, Marchante H, Marticorena A, Minter DW, Moreno RA, Nentwig W, Niamir A, Novoa A, Nunes AL, Pauchard A, Rahlao S, Ricciardi A, Russell JC, Sankaran KV, Schertler A, Schwindt E, Shackleton RT, Simberloff D, Strayer DL, Tawake A, Thines M, Villaseñor-Parada C, Vitule JRS, Wagner V, Werenkraut V, Wesche K, Willette DA, Zenni RD, Pyšek P (2025) Biological invasions: a global assessment of geographic distributions, long-term trends, and data gaps. Biological Reviews 100: 2542–2583. https://doi.org/10.1111/brv.70058

Vanderhoeven S, Adriaens T, Desmet P, Strubbe D, Backeljau T, Barbier Y, Brosens D, Cigar J, Coupremanne M, Troch RD, Eggermont H, Heughebaert A, Hostens K, Huybrechts P,

Jacquemart A-L, Lens L, Monty A, Paquet J-Y, Prévot C, Robertson T, Termonia P, Kerchove RVD, Hoey GV, Schaeybroeck BV, Vercayie D, Verleye TJ, Welby S, Groom QJ (2017) Tracking Invasive Alien Species (TrIAS): Building a data-driven framework to inform policy. Research Ideas and Outcomes 3: e13414. https://doi.org/10.3897/rio.3.e13414

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLOS ONE 7: e29715. https://doi.org/10.1371/journal.pone.0029715

Additional information

Conflict of interest

The authors have declared that no competing interests exist.

Ethical statement

No ethical statement was reported.

Artificial Intelligence (AI) use

The authors accept full responsibility for the content of the manuscript, including the disclosure of any use of AI.
Regarding the use of AI in the preparation of this manuscript, the authors declare the following: Claude.
Used for: Other.
Authors used AI for assistance with code optimization (software and figures).

Funding

The funding to TRT and HS for this project is provided by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), Kennedyallee 40, 53175 Bonn, Germany: "PredictiveAliens-1" (project no. 521529463) and "RobustTrends" (project no. 534019950). FE appreciates funding by the Austrian Science Fund FWF (pr. no. FW77206 and I 5825-B).

Author contributions

T.R.-T.: analysis, writing – original draft. H.S.: supervision, writing – review & editing. F.E.: writing – review & editing. All authors: conceptualization, methodology, data curation, read and approved the final manuscript.

Author ORCIDs

T. Renard Truong https://orcid.org/0009-0008-4580-7154
F. Essl https://orcid.org/0000-0001-8253-2112
H. Seebens https://orcid.org/0000-0001-8993-6419

Data availability

All of the data that support the findings of this study are available in the main text or Supplementary Information.

Supplementary material 1

Supplementary tables

Authors: Tanara Renard Truong, Franz Essl, Hanno Seebens

Data type: xlsx

Explanation note: **table S1.** Fields of the First Record main dataset. **table S2.** Fields of the Taxonomy table. **table S3.** Fields of the Location table. **table S4.** Fields of the Source table. **table S5.** Sources of the FirstRecords dataset. **table S6.** Sample FirstRecords dataset.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/neobiota.110.198728.suppl1