NeoBiota 59: 39–59 (2020)
doi: 10.3897/neobiota.59.53578
http://neobiota.pensoft.net

SOFTWARE DESCRIPTION

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

A workflow for standardising and integrating alien species distribution data

Hanno Seebens[1], David A. Clarke[2], Quentin Groom[3], John R. U. Wilson[4,5], Emili García-Berthou[6], Ingolf Kühn[7,8,9], Mariona Roigé[10], Shyama Pagad[11], Franz Essl[12], Joana Vicente[13], Marten Winter[9], Melodie McGeoch[2]

1 *Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25, 60325 Frankfurt, Germany* **2** *School of Biological Sciences, Monash University, Clayton 3800, VIC, Australia* **3** *Meise Botanic Garden, Meise, Belgium* **4** *Centre for Invasion Biology, Department of Botany and Zoology, Stellenbosch University, South Africa* **5** *South African National Biodiversity Institute, Kirstenbosch Research Centre, Cape Town, South Africa* **6** *GRECO, Institute of Aquatic Ecology, University of Girona, 17003 Girona, Spain* **7** *Helmholtz Centre for Environmental Research – UFZ, Department of Community Ecology, Theodor-Lieser-Str. 4, 06120 Halle, Germany* **8** *Martin Luther University Halle-Wittenberg, Geobotany and Botanical Garden, Am Kirchweg 2, 06108 Halle, Germany* **9** *German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany* **10** *AgResearch, Biocontrol and Biosecurity, Private Bag 4749, Christchurch 8140, New Zealand* **11** *IUCN Species Survival Commission Invasive Species Specialist Group (ISSG), University of Auckland, Auckland 1072, New Zealand* **12** *Department of Botany and Biodiversity Research, University Vienna, Rennweg 14, 1030 Vienna, Austria* **13** *Research Centre in Biodiversity and Genetic Resources (CIBIO) / InBIO Research Network in Biodiversity and Evolutionary Biology, Campus Agrário de Vairão, Rua Padre Armando Quintas nº 7, 4485-641, Vairão, Vila do Conde, Portugal*

Corresponding author: Hanno Seebens (hanno.seebens@senckenberg.de)

Academic editor: Maud Bernard-Verdier | Received 23 April 2020 | Accepted 3 July 2020 | Published 28 July 2020

Citation: 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

Abstract

Biodiversity data are being collected at unprecedented rates. Such data often have significant value for purposes beyond the initial reason for which they were collected, particularly when they are combined and collated with other data sources. In the field of invasion ecology, however, integrating data represents a major challenge due to the notorious lack of standardisation of terminologies and categorisations, and the application of deviating concepts of biological invasions. Here, we introduce the SInAS workflow, short for Standardising and Integrating Alien Species data. The SInAS workflow standardises terminologies following Darwin Core, location names using a proposed translation table, taxon names based on the GBIF backbone taxonomy, and dates of first records based on a set of predefined rules. The output of the SInAS workflow provides various entry points that can be used both to improve coherence among the databases and to check and correct the original data. The workflow is flexible and can be easily adapted and extended to the needs of different users. We illustrate the workflow using a case-study integrating five widely used global databases of information on biological invasions. The comparison of the standardised databases revealed a surprisingly low degree of overlap, which indicates that the amount of data may currently not be fully exploited in the original databases. We highly recommend the use and development of publicly available workflows to ensure that the integration of databases is reproducible and transparent. Workflows, such as SInAS, ultimately increase trust in data, study results, and conclusions.

Keywords
databases, Darwin Core, GBIF, invasive alien species, R software environment, reproducibility, standardisation, taxonomy, workflow

Introduction

In recent years, we have observed a tremendous rise in the availability of data in all fields of biodiversity research (La Salle et al. 2016), including invasion ecology. In particular, initiatives have emerged to map the occurrence of specific taxa with alien populations – called ‘alien taxa’ in the following – for major groups such as plants, birds, amphibians and reptiles (van Kleunen et al. 2015; Dyer et al. 2017a; Capinha et al. 2017); to assess the extent of invasions in particular geographical regions (e.g., Europe, DAISIE 2009) and habitats (e.g., marine, Ahyong et al. 2019); to document particular events (e.g., dates of record, Seebens et al. 2017); or to identify and record the presence of alien species that have negative impacts (e.g., Pagad et al. 2018). Although analyses of these data sources have led to valuable insights on the historic and current spatial and temporal patterns and processes of biological invasions (Dyer et al. 2017a; Dawson et al. 2017; Pyšek et al. 2017; Bertelsmeier et al. 2017; Seebens et al. 2018), these new aggregations of alien species data differ in various respects and are not interoperable.

Biodiversity data sources are often not standardised or directly comparable (Guralnick et al. 2018), which limits their value for conservation and research (Bayraktarov et al. 2019). In invasion ecology, new databases have recently been produced for a range of different purposes, although they have, to date, been produced largely in isolation. To remedy this, individual workflows have been created to harmonise and integrate the information in order to meet particular project goals. These workflows have used different taxonomic and geographical standards and practices, but such standardisations are not always clearly documented. As a result, databases are often not comparable and cannot be readily linked, which hampers progress towards improving the taxonomic and geographic coverage of alien species data and potential insights for research and management that might be derived as a consequence (McGeoch et al. 2012). The widespread lack of standardisation across key data sources on alien species also hinders clear communication with managers and policy makers (Gatto et al. 2013; McGeoch and Jetz 2019).

Progress in biodiversity research has been facilitated by the development of data standards (Guralnick and Hill 2009), powerful analytical tools and coherent workflows to, for instance, develop and calculate Essential Biodiversity Variables (EBVs, Kissling et al. 2018; Jetz et al. 2019) or to clean biodiversity data (Mathew et al. 2014; Jin and Yang 2020). Recently, using three exemplar alien species, a workflow was constructed and tested to integrate data from multiple sources for alien species (Hardisty et al. 2019). For most comprehensive databases in invasion ecology, the publication of such workflows and detailed descriptions of database generation remains rare (but see Dyer et al. 2017b; Pagad et al. 2018). Thus, data management in invasion ecology does not often meet open science principles, and the databases produced do not qualify as FAIR, i.e. Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). Although the procedures for collating data are often described, the descriptions and associated metadata are generally insufficient for the workflow to be reproduced. Computer scripts and guidance documents are often not publicly available, which further impedes reproducibility. Using a standardised, publicly available workflow would enable alien species databases to be combined in a transparent and repeatable way, and improve the format, contents, and interoperability of databases (Mathew et al. 2014). Such annotated workflows would also guide future data collation efforts such that they achieve both their own goals and contribute to community-wide efforts to enhance the quality and quantity of data on alien and invasive species (Hobern et al. 2019). In particular, any integration of species databases requires a well-documented, repeatable, coherent, and standardised workflow to match nomenclature and taxonomy based on a standard concept (e.g., Boyle et al. 2013; Murray et al. 2017), or even to map different taxonomic concepts to each other (Berendsohn 1995). The availability of large online infrastructures for biodiversity research, such as the Global Biodiversity Information Facility (GBIF), enables taxonomic standardisation in a reproducible and standardised way, but the potential is still not fully exploited in studies addressing biological invasions.

Here, we introduce the SInAS (Standardising and Integrating Alien Species data) workflow that was developed within the course of the synthesis working group "Theory and Workflows for Alien and Invasive Species Tracking" (sTWIST) at sDiv, Leipzig, Germany. Following Hardisty and Roberts (2013), we use the term "workflow" as a description of a series of processes of data manipulation and integration, including the codes allowing a largely automated approach (see also van der Aalst and van Hee 2002, who use the term "workflow" for a series of standardised processes). The SInAS workflow serves to integrate databases of regional checklists including information on spatial and temporal dynamics of alien species using a standardised protocol to merge taxon and location names. The SInAS workflow combines public taxonomic infrastructures with procedures, resolutions, and concepts commonly used in biodiversity research in general and invasion ecology in particular. In the following, we provide a detailed description of the SInAS workflow and its implementation in R. We demonstrate its functionality using an example of merging five of the most comprehensive open access alien species databases currently available. Although the workflow was developed for merging databases of alien species occurrences, it can be readily adapted to other databases, including those with associated spatial information.

The SInAS workflow

The SInAS workflow was created to integrate databases organised as individual spreadsheet tables, which is the most common format for alien species occurrence information. In contrast to databases of native species, alien species occurrences are often associated with a date of first introduction or first date of report for a region as an alien or naturalised species. Here, we adopt a common use of these "first records", which represent the first record of a taxon in a particular region. Following Darwin Core terminology (Darwin Core Task Group 2009), first records are called "event dates" in the following.

Three major steps, organised in sequence, form the primary components of the workflow: 1) initial check and preparation of the original databases; 2) standardisation of the databases; and 3) merging of the standardised databases (Fig. 1). Standardisation (step 2) is the most complex step and can be subdivided into specific tasks that each involves the standardisation of one of eight variables: taxon names, location names, event dates, occurrence status, establishment means, degree of establishment, pathway, and habitat. An overview of all variables used in this workflow together with definitions and explanations are given in Suppl. material 2: Tables S1–S4. Each specific task requires a reference against which data will be standardised (e.g., a list of location names in a particular format or a list of accepted taxon names and their synonyms). Each task produces intermediate output tables to report where there was standardisation (e.g., replacements of original names) and where standardisation was not possible (e.g., missing names and unresolved names). As input files, each step of the workflow requires the output of the previous step as input except for step one, where the original database and its metadata have to be provided (currently implemented as *.xlsx files). In the following section, a comprehensive overview of the SInAS workflow is provided, while the detailed description can be found in the Suppl. material 1. The full workflow implemented in R together with all required input files, examples databases, and a manual are provided as the SInAS workflow package (see section 'Data and code availability' below).

Step 1: Preparation of databases

The first step includes a check of the availability of variables in the original databases. Variables are categorised into three classes: i) required variables, which must be provided (i.e., taxon and location names); ii) optional variables, which are associated to the taxon occurrence (e.g., occurrence status or pathway) or represent entries potentially useful for data standardisation (e.g., extra taxonomic information); and iii) additional variables, which are not used within the workflow, but are retained as presented in the original databases throughout standardisation (e.g., traits). An overview of variables and definitions is provided in Suppl. material 2: Table S1. The column names of the required and optional variables in the input databases are harmonised.

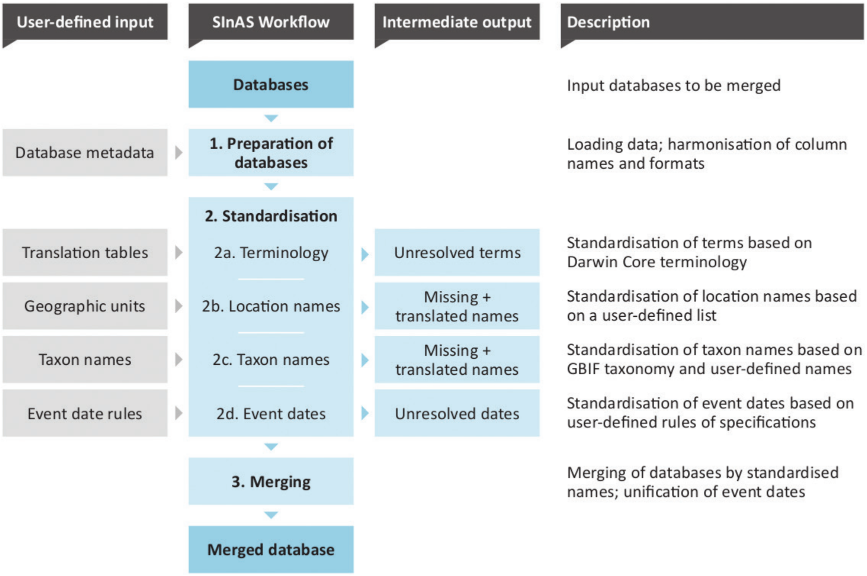
User-defined input
SInAS Workflow
Intermediate output
Description
Databases
Input databases to be merged
Database metadata
1. Preparation of databases
Loading data; harmonisation of column names and formats
2. Standardisation
Translation tables
2a. Terminology
Unresolved terms
Standardisation of terms based on Darwin Core terminology
Geographic units
2b. Location names
Missing + translated names
Standardisation of location names based on a user-defined list
Taxon names
2c. Taxon names
Missing + translated names
Standardisation of taxon names based on GBIF taxonomy and user-defined names
Event date rules
2d. Event dates
Unresolved dates
Standardisation of event dates based on user-defined rules of specifications
3. Merging
Merging of databases by standardised names; unification of event dates
Merged database

Figure 1. Overview of the Standardising and Integrating Alien Species data (SInAS) workflow that can be used to merge alien species databases. The workflow consists of three consecutive steps: 1. preparation of databases, 2. standardisation, and 3. merging. The standardisation step is subdivided into the standardisation of: 2a. terminology, 2b. location names, 2c. taxon names, and 2d. event dates (i.e., first records). The user can modify the workflow by adjusting the reference tables under 'user-defined input'. At each step of standardisation, changes and missing entries are exported as intermediate output that can be used to check the workflow, the reference tables, or the input data.

Step 2: Standardisation

2a: Terminology

Records of alien species are often associated with information about their occurrence status, the degree of establishment, and their pathway(s) of introduction. Such information is standardised in this step using translation tables (Suppl. material 1). Translation tables provide information about the entries in the original databases and the corresponding terms that are to be used in the merged database. These are part of the workflow package (see section 'Data and code availability' below), and follow the recommendations by Groom et al. (2019) in standardising the Darwin Core terms 'establishmentMeans', 'occurrenceStatus' and 'pathway', and adopting their suggestion to include a new term 'degreeOfEstablishment', describing the status of the taxon at a particular location (Suppl. material 2: Table S1). Strictly speaking, this status is not associated to a taxon, but a specific population. This means, as Colautti & MacIsaac (2004) already pointed out, that alien or nonindigenous species are misnomers and these attributes, frequently referred to simply as "status", are associated at population level (i.e., intersecting taxon name with locality). In databases covering large regions, such attributes must properly be assigned at the right level. However, to be comparable with the wealth of invasion literature that does not properly attribute "status", and for reasons of linguistic simplicity, we still refer to alien species rather than using the correct alien populations. Although the proposal by Groom et al. (2019) has not yet been ratified by the Biodiversity Information Standards organisation, we used it in the workflow as the proposed terminology covers dimensions critical to invasion biology, policy, and management (McGeoch and Jetz 2019), and thus will provide helpful information irrespective of its official incorporation into Darwin Core. The Darwin Core term 'habitat' is also standardised within the workflow; however, as a categorisation of different habitats is not provided by Darwin Core, we provide one in the respective translation table (Suppl. material 1) based on the distinction between terrestrial, freshwater, marine, and brackish habitats. The translation tables can be adjusted by the user in any way, but we highly recommend adhering to the proposed Darwin Core terminology to avoid having incomparable entries. Non-matching terms are exported so they can be manually checked.

2b: Location names

Location names are standardised using a user-defined translation table (Suppl. material 1), which includes the master location names and the corresponding alternative formats, languages, and spellings. Locations represent administrative units such as countries, states or islands. The majority of location names (89%) conform to the 2-digit ISO code (ISO 3166-1 alpha-2) classification. For the remaining locations, countries were split into sub-national units which are geographically separated from each other (be they islands, states or mainland areas). For instance, Alaska, Hawaii, and US Minor Outlying Islands were separated from mainland United States; the Azores were distinguished from Portugal; and Tasmania from Australia. The full list of location names can be found in the input file "AllLocations.xlsx" as part of the workflow package. Altogether, we used a set of 262 non-overlapping locations covering the terrestrial surface of the world. Similar resolutions are used in many studies of biological invasions (Seebens et al. 2017; Capinha et al. 2017; Dyer et al. 2017b). The location categorisation can be easily adjusted to any spatial delineation in a user-friendly way by modifying the input file. Additional information for the location such as two- and three-digit ISO codes of countries, continents or the World Geographical Scheme for Recording Plant Distributions regions (WGSRPD, Brummitt 2001) are also provided. Non-matching location names are exported for reference. A shapefile is provided, which relates the location to georeferenced polygons for mapping.

2c: Taxon names

Taxonomic standardisation is one of the most important and challenging tasks in biodiversity data integration (Rees and Cranston 2017) as taxon names are often considered the fundamental unit to which other information types are linked (Patterson et al. 2010; Koch et al. 2018). This, however, necessitates the use of a taxonomic backbone against which all species names are assessed during the standardisation process. In the absence of a single authoritative nomenclature across all taxa (Bánki et al. 2018), we used the GBIF taxonomic backbone, which is itself primarily based on the Catalogue of Life (Bánki et al. 2018) (43 % overlap of GBIF backbone taxonomy and Catalogue of Life at the time of access) and complemented with 50+ other sources of taxonomic information. The details of these taxonomic sources can be found at the GBIF Secretariat (2019) and the full taxonomy is available for download (http://rs.gbif.org/datasets/backbone/). If the taxon name could be found in GBIF either as an exact match, a synonym or a fuzzy match with a high confidence (see Suppl. material 1), the obtained 'accepted taxon name' according to GBIF, as well as its given synonym and further taxonomic information, are returned and stored. Taxon names identified as synonyms according to GBIF are replaced with the accepted name obtained from GBIF. To avoid mismatches due to spelling errors, GBIF performs fuzzy matching of the full taxon names. This involves a calculation of similarity between the provided taxon names and the record provided by GBIF. GBIF returns the result of fuzzy matching by the summary metric "confidence", which involves cross-checks of taxon names, authorities and taxonomic information with different weightings (see http://www.gbif.org/developer/species#searching for more details). In addition to the taxon names, the taxonomic tree (species, genus, family, order, class, phylum, and kingdom) is obtained from GBIF. In the SInAS workflow, all taxon names that could not be resolved are exported as a list of missing taxon names for further reference. A complete list of all taxon names (including the original names provided in the individual databases, taxonomic information, taxonomic status of the name, and search results) is exported as a separate list of taxon names (Suppl. material 1). The user can provide a list of species names and synonyms to resolve conflicts and errors in GBIF entries.

2d: Event dates

In the SInAS workflow presented here, event dates represent the time of the first documented occurrence of a species in a region outside its native range, which is also called 'first record' (Seebens et al. 2017). Ideally, event dates for the first record of an alien species are provided as a single year, which is then retained in the workflow. But often other time ranges are provided. To enable merging and cross-checking of first records among databases and further analysis, it is necessary to translate these different time ranges into single years. Such an adjustment of first records requires a set of rules (e.g., Seebens et al. 2017; Dyer et al. 2017b), which define how a time range should be treated to obtain a single year. In the simplest case, the start and the end years of the time range are provided, and their arithmetic mean is used as the new single event date. In other cases, time ranges are described in alternative ways such as "1920ies" or "<1920". In translating this information, we followed primarily the rules defined in table 3 of Dyer et al. (2017b). The rules are currently provided as a textual description and the user has to "translate" non-standard event dates into a single year format according to the guidelines and examples provided in the file 'Guidelines_eventDate.xlsx' as part of the workflow package. The user has the opportunity to modify the rules, but we recommend sticking to the proposed ones as a standard in biological invasions. Cases of entries that could not be adjusted are exported from the workflow for cross-checking.

Step 3: Merging

In the final step of the workflow, the standardised databases are merged into a single master database. Merging is based on the entries of taxon and location names. That is, all entries with exactly the same taxon and location name will be merged to obtain a single entry for each existing combination of taxon and location. This is achieved by first merging columns of the standardised databases to concatenate their contents and, second, by merging rows of the final database to remove duplicate entries. Conflicts of multiple event dates for the same event are resolved by adopting the earlier of the first records. In cases where conflicts cannot be resolved, the respective entries of all databases are combined to a single entry of the master database. For instance, if a taxon X in location Y is classified as 'introduced' in one database and 'uncertain' in another, the entry in the final master database for X in Y will be 'introduced; uncertain'. The user will be informed that conflicts still exist, which might be solved by adjusting the translation tables or by checking the original data.

In principle, the SInAS workflow is fully automated once metadata are provided at step 1. This, however, requires accepting all defaults such as location names and taxonomic classification by GBIF and, more importantly, keeping all unresolved conflicts that might include unmatched location names or misspellings in the original data. We therefore recommend running the workflow in an iterative process of running the workflow, checking warnings and intermediate output tables, resolving conflicts and errors, and re-running the workflow. Such an iterative process should increase the match between databases, and therefore the coverage of the final merged database.

A case study

We applied and tested the workflow using five global databases of spatio-temporal alien species occurrences (Table 1): three with a taxonomic focus, one each on alien birds (GAVIA, Dyer et al. 2017b), vascular plants (GloNAF, van Kleunen et al. 2019), and amphibians and reptiles (AmphRep, Capinha et al. 2017); one multi-taxon database with a focus on temporal dynamics (FirstRecords version 1.2, Seebens et al. 2017); and one with a focus on alien species with negative environmental or socio-economic impacts, i.e. "invasive alien species" (GRIIS, Pagad et al. 2018; accessed 10th September 2019). These databases are currently among the most up-to-date and comprehensive global data sources for alien species distributions, dynamics, and impacts. All databases are publicly available. The lack of accessibility impeded the incorporation of other global databases such as the World Register of Marine Introduced Species (WRiMS) or the CABI Invasive Species Compendium. The databases used here are of varying size, ranging from 1,118 (AmphRep) to 232,042 (GloNAF) records and including 277 (AmphRep) to 33,687 (GRIIS) taxa. The databases have different spatial resolutions and follow different taxonomic standards. Variables from the different databases were mapped onto the variables provided in the SInAS workflow as outlined in Suppl. material 2: Tables S1–S4. As location names were

Table 1. The taxonomic coverage and size of the original databases on the occurrence of alien taxa before and after standardisation and merging using the Standardising and Integrating Alien Species data (SInAS) workflow (see Figure 1). Records were counted multiple times when they were obtained from different databases. Reductions in total record number were mostly a result of aggregation from the finer spatial resolution of the original databases to the higher spatial resolution used in the SInAS workflow.

Database	Reference	Focus of database	Total records		Number of taxa	
			(original)	(merged)	(original)	(merged)
GloNAF	van Kleunen et al. (2019)	Vascular plants	232,042	71,468	14,053	13,545
AmphRep	Capinha et al. (2017)	Amphibians, reptiles	1,118	854	277	276
GAVIA	Dyer et al. (2017b)	Birds	27,723	4,494	971	968
GRIIS	Pagad et al. (2018)	Invasive species	107,302	96,655	33,687	27,128
FirstRecords	Seebens et al. (2017)	First records	45,402	45,060	15,231	14,990

provided in different columns in GloNAF and GAVIA, these were merged manually to obtain a better match with the classification of locations used in the SInAS workflow.

Merging of the five databases resulted in a new database (the sTWIST database) consisting of two interlinked tables containing records of alien species per location and a full list of taxa including further taxonomic information (Suppl. material 3). Depending on the success of the integration of the specific databases, several additional files will be created during the workflow providing missing taxa and location names, unresolved terms (e.g., of occurrence status and pathways), translated location names and event dates, and unresolved event dates. In our cases, 17 of these tables were exported from the workflow for further cross-checking (Suppl. material 5) together with 25 tables, which include the output of each intermediate step and database (Suppl. material 4). The sTWIST database contains 156,900 records of 35,150 taxa in 257 locations. The resulting alien species numbers globally are in line with the reported hotspots of biological invasions being the USA (excluding Hawaii and Alaska), the United Kingdom, New Zealand, Hawaii, and Australia (fig. 2, Dawson et al. 2017). One consequence of the workflow was that, after cleaning and standardisation, the number of records dropped (Table 1). For example, the merged sTWIST database contained only ~30% of the original GloNAF database. This was mostly due to the GloNAF database having a finer spatial resolution than the sTWIST database (1,029 vs. 257 regions). Consequently, many regions were combined and records merged.

Altogether, 53,546 taxon names were obtained from all five databases, including synonyms and multiple entries of individual taxa due to different spellings. A small proportion (5 %) of these taxon names could not be found in GBIF for different reasons such as misspellings, missing information or unresolved taxonomies. This often involved subspecies, varieties or hybrids and can be checked in the output files "Missing_Taxa_*" for the individual databases. Most of these unresolved taxon names were obtained from GRIIS (1,610; 6 % of GRIIS taxa) followed by FirstRecords (802; 5%), AmphRep (10; 4%), GloNAF (261; 2%) and GAVIA (8; <1%). Unresolved taxon names were kept in the final database but flagged as such in the full list of taxon names "Taxa_FullList.csv". Standardisation during the SInAS workflow identified 7,174 syn-

Table 2. Overlap (in %) of locations, taxa, and taxa by location record between taxonomic and cross-taxon databases. An overlap between two databases is defined as the number of entries in the taxon-specific database shared with the cross-taxon database divided by the total number of entries from the taxon-specific database. It therefore shows how many records of the taxon-specific databases are found in the cross-taxon ones.

	GRIIS	FirstRecords
Locations		
GloNAF	76	97
GAVIA	76	98
AmphRep	74	98
Taxa		
GloNAF	69	45
GAVIA	54	86
AmphRep	61	63
Taxa by location		
GloNAF	44	20
GAVIA	26	78
AmphRep	29	41

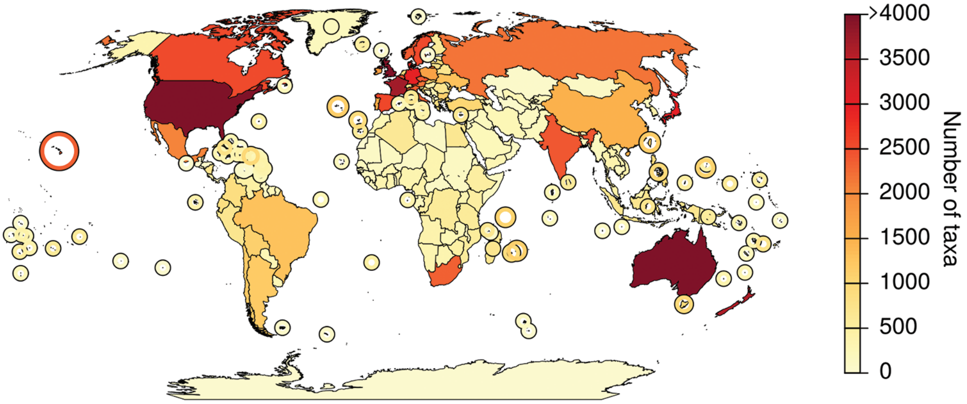

Figure 2. The number of alien taxa per region as presented in the final sTWIST database. Smaller island regions are depicted by circles, with the size of the circles proportional to the numbers of taxa. Region delineations are based on Global Administrative Areas (GADM).

onyms (13%), which were replaced by the accepted names provided by GBIF. This finally reduced the number of taxa to 35,150 distinct taxon names.

After standardisation of taxon and location names, the overlap of taxon-specific databases with the cross-taxon ones was surprisingly low (Table 2). Most regions were represented in all databases; however, the overlaps for taxa and taxon by location combinations were often far below 50%. For instance, only 26% of all records in GAVIA can also be found in GRIIS, while 20% of the GloNAF records were also included in FirstRecords. The comparatively low overlap of locations in GRIIS with taxon-specific databases stems from a few locations only considered separately in GRIIS.

Discussion

The SInAS workflow is, to the best of our knowledge, the most comprehensive workflow to standardise and integrate alien species occurrence databases to date. It is also in full compliance with the FAIR data principles (Wilkinson et al. 2016). The workflow provides a foundation to develop and apply standards for the harmonisation of taxon names, geographic resolutions, and event dates. It achieves this using translation tables and rules that are transparent and linked to existing international schemes such as accepted taxonomic backbones that can be easily updated as needed. The SInAS workflow also offers the opportunity to adapt individual steps to the respective user's needs, and enables the user to conveniently report on deviations from the suggested workflow. Reporting of such adjustments is essential for reproducibility, particularly in the field of invasion ecology, which is rich in competing concepts and terminologies (Falk-Petersen et al. 2006). Thus, the SInAS workflow will help to differentiate and integrate the various approaches, and finally will increase trust not only in data but also in study results and conclusions communicated to the decision makers and the general public (Franz and Sterner 2018). The potential to customise and extend the workflow increases the range of possible applications such as the calculation of indicators (e.g., Wilson et al. 2018), the ability to conduct global and regional assessments of invasive alien species and their control, and the global collaboration being proposed as essential for dealing with priority invaders (Blackburn et al. 2020).

We introduced the SInAS workflow as a tool to integrate databases, but it can also assist with standardisation within a database to ensure that region or taxon names are consistent, and that terminologies of individual checklists are reported in a more standardised way. Although the flexibility built into the SInAS workflow makes it more broadly useful, providing flexibility in a workflow does bear the risk that databases remain incompatible. For instance, users of the workflow can define their own categorisation of locations, which might result in even more heterogeneous databases in addition to those that already exist. It is essential, therefore, that modifications of the workflow are clearly communicated. As best practice, we recommend that modifications of the input files such as translation tables, taxon names or any modification of the workflow itself are clearly reported and published together with the final database. For instance, a change in the list of geographic regions can be easily attached as a table to the respective publication together with the link to our workflow. In this way, modifications can be traced back to their origin and databases remain comparable despite adaptations to individual project goals. We believe that our proposed workflow will smooth this process and make it easier for individual researchers to publish not only scientific results in a more consistent way, but also the underlying workflows to enhance the transparency and reproducibility of the science.

The comparison of the individual databases that resulted from the integration work done here highlighted an unexpectedly low degree of overlap between them. This re-emphasizes, in spite of significant recent advances in alien species data collation, the importance of: 1) joint collaborative work, 2) freely available data, and 3) shared workflows to improve the taxonomic, geographic, and temporal coverage and resolution of alien species data (Hardisty et al. 2019). The low degree of overlap was obviously related to the scope of the individual databases – the taxon-specific databases focussed on a high level of spatial and taxonomic coverage, while cross-taxonomic databases harvest information on a specific topic such as event dates or impact. Moreover, the databases drew original data records from different sources, and so each database was constructed using different workflows with divergent assumptions and supporting concepts. This clearly shows that not only does the merging of individual databases have to be standardised as proposed here, but the integration of primary data from the original sources needs to be done in a more reproducible and transparent way as well (Vanderhoeven et al. 2017; Pagad et al. 2018). Our case study also highlights that the SInAS workflow and the associated scripts could be used to assess the reliability of different databases and their components (e.g., Cano-Barbacil et al. 2020) and to identify potential areas of improvement for the respective databases.

Our workflow was developed to integrate taxon lists for individual regions, so-called checklists. Checklists represent by far the most common representation of spatial information on alien species occurrences (Pyšek et al. 2012; Brundu and Camarda 2013). This is somewhat different to other fields of biodiversity research, where occurrence data are often provided as range maps, grids, plot based lists or point coordinates. In contrast to populations of native taxa, alien taxa populations are categorised as being alien only for a particular region and timeframe. The importance of decision-making in an applied science, such as invasion ecology, means that policies are commonly made for the administrative units (such as countries or states/provinces) responsible for control efforts, and the spatial resolution of presence-absence data is low resolution to accommodate both uncertainty and the precautionary principle when data are intended to inform policy and management. As a consequence, the decision of what is considered as being alien is often taken for administrative regions. This is somewhat different for aquatic alien species, which are categorised depending on marine regions or water sheds, but these spatial units can be easily incorporated as additional entries in the table of geographic regions. In its current form, the SInAS workflow is not capable of handling coordinate-based occurrences. While including point-wise occurrences might be possible in future versions of the workflow, a practical solution would be to assign the coordinate-based location to a region and add the region to the workflow. For example, point-wise occurrence data for the Western Mediterranean Sea could be attributed to this region and added to the workflow.

The pervasive challenge in the integration of alien species data from multiple sources is the variability in the use of terminology (McGeoch et al. 2012). For example, the term ‘invasive species’ has at least three working definitions: alien populations that are self-sustaining and have naturally spread; alien populations that negatively impact native species, ecosystems, the economy or human health; or populations (be they native or alien) that have recently increased in abundance or extent (Richardson et al. 2000; Blackburn et al. 2011; Carey et al. 2012). As a consequence, merging databases that use different definitions of alien and invasive alien species could result in a misleading collation of taxa. Currently, terminologies are not consistently used across databases, although standard concepts have been published (Blackburn et al. 2011). In the SInAS workflow, we provide a translation of terms following common standards (Darwin Core Task Group 2009; Groom et al. 2019), but the definitions of these terms may vary among primary sources and projects, which often cannot be standardised retrospectively. It is therefore essential to stick to common definitions and transparent workflows already in the primary literature, to clearly specify which definition is used.

A further difficulty in combining species data lies in the application of different taxonomic concepts (Berendsohn 1995) by the data recorders. This is a general problem in biodiversity and taxonomic research and is not solved within the SInAS workflow: it requires collaborative solutions from the relevant research community. While resolving such taxonomic conflicts would mean the SInAS workflow is more useful, one should keep in mind that a complete taxonomic resolution is not necessarily required to provide useful information (Gerwing et al. 2020). Unless this workflow is used by experienced taxonomists for taxonomic resolution, we recommend sticking to standards offered by other authorities such as GBIF and report deviations from these standards. Our workflow eases this reporting process by providing the opportunity to submit information of modifications together with the databases.

While advancements have been made in other fields of biodiversity research, with online platforms such as GBIF including a full and citable version control, many databases on biological invasions are still curated by individuals or research groups and might not be publicly available at all. Changing this situation will require there being: 1) an incentive for researchers to publish their data online, ideally with a digital object identifier (DOI) and versioning as provided by online platforms such as GBIF or long-term archives such as Zenodo (https://zenodo.org/) or Dryad (https://datadryad.org), and following the FAIR principles of data management; 2) professional training and technical support for data management; and 3) clear guidelines and standards to ease such data publications (Groom et al. 2019). For some of these aspects, support is already available but still not widely adopted such as the “Guide to Data Management in Ecology and Evolution” published by the British Ecological Society (2014). For other aspects, financial and personnel support is required as individual researchers often do not have the capacity to ensure long-term maintenance and support, which can only be achieved from institutions. The importance of adopting the FAIR data principles has been increasingly recognised by international institutions such as the Intergovernmental Science-Policy Platform of Biodiversity and Ecosystem Services [IPBES, currently conducting a thematic assessment on invasive alien species and their control (https://ipbes.net/invasive-alien-species-assessment) that depends on the integration of data sources as we have discussed here] and the European Commission, which provide incentives to scientists to make their data comparable and available. We believe the workflow presented here addresses these challenges by providing an example of how to achieve standardisation across databases and to facilitate the kind of standardisation chosen by the researchers.

The modular structure of the SInAS workflow means that it can form the basis for the development of future data integration workflows. We foresee several opportunities for extensions. Translation tables of additional variables such as taxon traits and variables related to regions and relevant for understanding drivers of biological invasions (environmental, socio-economic, historic) would add another level of value for both research and application. The workflow could also be extended to allow for coordinate-based occurrence records by integrating information of region delineations using Geographic Information System (GIS) tools. Thus, the SInAS workflow, focussed as it is on essential variables for tracking biological invasions (distribution, time, and impact, Latombe et al. 2017), can be considered the core of an integrated comprehensive workflow of data on biological invasions. Global collaborative efforts, supported by readily accessible, globally representative evidence, are key to stemming the invasion tide.

Data and code availability

The full SInAS workflow including all required R scripts, input files, example databases and a manual is made freely available at a repository at Zenodo (https://doi.org/10.5281/zenodo.3944432) together with the coordinate-based delineations of regions. The releases at Zenodo are linked to a GitHub repository, which ensures full version control of the code. New releases will be provided under the same DOI. All additional files related to the case study are attached to this publication as supplementary materials.

Acknowledgements

This paper is a joint effort of the sTWIST working group (Theory and Workflows for Invasive Species Tracking) supported by sDiv, the Synthesis Centre of iDiv (DFG FZT 118 – 202548816). It is a contribution to the Species Populations Working Group of the Group on Earth Observations Biodiversity Observation Network (GEO BON; https://geobon.org/ebvs/workinggroups/species-populations). We thank Wolfgang Traylor for advice on structuring the R code, Carlos Eduardo Arlé Ribeiro de Souza for providing the shapefile and Gabriele Rada for support on graphic design. Support from the following funding agencies is acknowledged: HS – Belmont Forum-BiodivERsA project AlienScenarios through the national funders German Federal Ministry of Education and Research (BMBF; grant 01LC1807A). MAM – Australian Research Council (DP200101680). FE – BiodivERsA-Belmont Forum Project AlienScenarios (FWF project no I 4011-B32). JRUW – South African Department of Forestry, Fisheries and the Environment (DFFtE) for funding noting that this publication does not necessarily represent the views or opinions of DFFtE or its employees. DAC – Australian Government Research Training Program (RTP) scholarship. EGB – Spanish Ministry of Science and Innovation (projects CGL2016-80820-R, PCIN-2016-168 and RED2018-102571-T) and the Government of Catalonia (ref. 2017 SGR 548). QG – Belgian Science Policies Brain program (BR/165/A1/TrIAS).

References

Ahyong S, Costello MJ, Galil BS, et al. (2019) World Register of Introduced Marine Species (WRiMS).

Bánki O, Döring M, Holleman A, Addink W (2018) Catalogue of Life Plus: innovating the CoL systems as a foundation for a clearinghouse for names and taxonomy. Biodiversity Information Science and Standards 2: e26922. https://doi.org/10.3897/biss.2.26922

Bayraktarov E, Ehmke G, O'Connor J, et al. (2019) Do big unstructured biodiversity data mean more knowledge? Frontiers in Ecology and Evolution 7: 1–5. https://doi.org/10.3389/fevo.2019.00319

Berendsohn WG (1995) The concept of "potential taxa" in databases. Taxon 44: 207–212. https://doi.org/10.2307/1222443

Bertelsmeier C, Ollier S, Liebhold A, Keller L (2017) Recent human history governs global ant invasion dynamics. Nature Ecology & Evolution 1: 0184. https://doi.org/10.1038/s41559-017-0184

Blackburn GS, Bilodeau P, Cooke T, Cui M, Cusson M, Hamelin RC, Keena MA, Picq S, Roe AD, Shi J, Wu Y, Porth I (2020) An Applied Empirical Framework for Invasion Science: Confronting Biological Invasion Through Collaborative Research Aimed at Tool Production. Annals of the Entomological Society of America. https://doi.org/10.1093/aesa/saz072

Blackburn TM, Pyšek P, Bacher S, Carlton JT, Duncan RP, Jarošík V, Wilson JRU, Richardson DM (2011) A proposed unified framework for biological invasions. Trends in Ecology & Evolution 26: 333–339. https://doi.org/10.1016/j.tree.2011.03.023

Boyle B, Hopkins N, Lu Z, Garay JAR, Mozzherin D, Rees T, Matasci N, Narro ML, Piel WH, Mckay SJ, Lowry S, Freeland C, Peet RK, Enquist BJ (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics 14: 1–16. https://doi.org/10.1186/1471-2105-14-16

British Ecological Society (2014) A Guide to Data Management in Ecology and Evolution. https://www.britishecologicalsociety.org/wp-content/uploads/Publ_Data-Management-Booklet.pdf

Brummitt RK (2001) World Geographical Scheme for Recording Plant Distributions. Hunt Institute for Botanical Documentation, Carnegie Mellon University, Pittsburgh.

Brundu G, Camarda I (2013) The Flora of Chad: a checklist and brief analysis. PhytoKeys 23: 1–18. https://doi.org/10.3897/phytokeys.23.4752

Cano-Barbacil C, Radinger J, García-Berthou E (2020) Reliability analysis of fish traits reveals discrepancies among databases. Freshwater Biology 65: 863–877. https://doi.org/10.1111/fwb.13469

Capinha C, Seebens H, Cassey P, García-Díaz P, Lenzner B, Mang T, Moser D, Pyšek P, Rödder D, Scalera R, Winter M, Dullinger S, Essl F (2017) Diversity, biogeography and the global flows of alien amphibians and reptiles. Diversity and Distributions 23: 1313–1322. https://doi.org/10.1111/ddi.12617

Carey MP, Sanderson BL, Barnas KA, Olden JD (2012) Native invaders – challenges for science, management, policy, and society. Frontiers in Ecology and the Environment 10: 373–381. https://doi.org/10.1890/110060

Colautti RI, MacIsaac HJ (2004) A neutral terminology to define 'invasive' species. Diversity and Distributions 10: 135–141. https://doi.org/10.1111/j.1366-9516.2004.00061.x

DAISIE (2009) Handbook of Alien Species in Europe. Springer, Dordrecht.

Darwin Core Task Group (2009) Darwin Core (Kampmeier G, review manager) Biodiversity Information Standards (TDWG). http://www.tdwg.org/standards/450

Dawson W, Moser D, van Kleunen M, Kreft H, Pergl J, Pyšek P, Weigelt P, Winter M, Lenzner B, Blackburn TM, Dyer EE, Cassey P, Scrivens SL, Economo EP, Guénard B, Capinha C, Seebens H, García-Díaz P, Nentwig W, García-Berthou E, Casal C, Mandrak NE, Fuller P, Meyer C, Essl F (2017) Global hotspots and correlates of alien species richness across taxonomic groups. Nature Ecology & Evolution 1: 0186. https://doi.org/10.1038/s41559-017-0186

Dyer EE, Cassey P, Redding DW, Collen B, Franks V, Gaston KJ, Jones KE, Kark S, Orme CDL, Blackburn TM (2017a) The Global Distribution and Drivers of Alien Bird Species Richness. PLoS Biology 15: e2000942. https://doi.org/10.1371/journal.pbio.2000942

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

Falk-Petersen J, Bøhn T, Sandlund OT (2006) On the Numerous Concepts in Invasion Biology. Biological Invasions 8: 1409–1424. https://doi.org/10.1007/s10530-005-0710-6

Franz NM, Sterner BW (2018) To increase trust, change the social design behind aggregated biodiversity data. Database 2018: 1–12. https://doi.org/10.1093/database/bax100

Gatto F, Katsanevakis S, Vandekerkhove J, Zenetos A, Cardoso AC (2013) Evaluation of Online Information Sources on Alien Species in Europe: The Need of Harmonization and Integration. Environmental Management 51: 1137–1146. https://doi.org/10.1007/s00267-013-0042-8

GBIF Secretariat (2019) GBIF Backbone Taxonomy.

Gerwing TG, Cox K, Allen Gerwing AM, Campbell L, Macdonald T, Dudas SE, Juanes F (2020) Varying intertidal invertebrate taxonomic resolution does not influence ecological findings. Estuarine, Coastal and Shelf Science 232: 106516. https://doi.org/10.1016/j.ecss.2019.106516

Groom Q, Desmet P, Reyserhove L, Adriaens T, Oldoni D, Vanderhoeven S, Baskauf SJ, Chapman A, McGeoch M, Walls R, Wieczorek J, Wilson JRU, Zermoglio PFF, 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

Guralnick R, Hill A (2009) Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. Bioinformatics 25: 421–428. https://doi.org/10.1093/bioinformatics/btn659

Guralnick R, Walls R, Jetz W (2018) Humboldt Core – toward a standardized capture of biological inventories for biodiversity monitoring, modeling and assessment. Ecography 41: 713–725. https://doi.org/10.1111/ecog.02942

Hardisty A, Roberts D (2013) A decadal view of biodiversity informatics: challenges and priorities. BMC Ecology 13: 1–16. https://doi.org/10.1186/1472-6785-13-16

Hardisty AR, Belbin L, Hobern D, McGeoch MA, Pirzl R, Williams KJ, Kissling WD (2019) Research infrastructure challenges in preparing essential biodiversity variables data products for alien invasive species. Environmental Research Letters 14: 025005. https://doi.org/10.1088/1748-9326/aaf5db

Hobern D, Baptiste B, Copas K, Guralnick R, Hahn A, van Huis E, Kim E-S, McGeoch M, Naicker I, Navarro L, Noesgaard D, Price M, Rodrigues A, Schigel D, Sheffield CA, Wieczorek J (2019) Connecting data and expertise: a new alliance for biodiversity knowledge. Biodiversity Data Journal 7: e33679. https://doi.org/10.3897/BDJ.7.e33679

Jetz W, McGeoch MA, Guralnick R, Ferrier S, Beck J, Costello MJ, Fernandez M, Geller GA, Keil P, Merow C, Meyer C, Muller-Karger FE, Pereira HM, Regan EC, Schmeller DS, Turak E (2019) Essential biodiversity variables for mapping and monitoring species populations. Nature Ecology and Evolution 3: 539–551. https://doi.org/10.1038/s41559-019-0826-1

Jin J, Yang J (2020) BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. Global Ecology and Conservation 21: e00852. https://doi.org/10.1016/j.gecco.2019.e00852

Kissling WD, Ahumada JA, Bowser A, Fernandez M, Fernández N, García EA, Guralnick RP, Isaac NJB, Kelling S, Wouter L, McRae L, Mihoub J-B, Obst M, Santamaria M, Skidmore AK, Williams KJ, Donat A, Amariles D, Arvanitidis C, Bastin L, De Leo F, Willi E, Elith J, Hobern D, Martin D, Pereira HM, Pesole G, Peterseil J, Saarenmaa H, Schigel D, Schmeller DS, Segata N, Turak E, Uhlir PF, Wee B, Hardisty AR (2018) Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. Biological Reviews 93: 600–625. https://doi.org/10.1111/brv.12359

Koch MA, German DA, Kiefer M, Franzke A (2018) Database Taxonomics as Key to Modern Plant Biology. Trends in Plant Science 23: 4–6. https://doi.org/10.1016/j.tplants.2017.10.005

La Salle J, Williams KJ, Moritz C (2016) Biodiversity analysis in the digital era. Philosophical Transactions of the Royal Society. https://doi.org/10.1098/rstb.2015.0337

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

Mathew C, Güntsch A, Obst M, Vicario S, Haines R, Williams AR, de Jong Y, Goble C (2014) A semi-automated workflow for biodiversity data retrieval, cleaning, and quality control. Biodiversity Data Journal 2: e4221. https://doi.org/10.3897/BDJ.2.e4221

McGeoch M, Jetz W (2019) Measure and Reduce the Harm Caused by Biological Invasions. One Earth 1: 171–174. https://doi.org/10.1016/j.oneear.2019.10.003

McGeoch MA, Spear D, Kleynhans EJ, Marais E (2012) Uncertainty in invasive alien species listing. Ecological Applications 22: 959–971. https://doi.org/10.1890/11-1252.1

Murray BR, Martin LJ, Phillips ML, Pyšek P (2017) Taxonomic perils and pitfalls of dataset assembly in ecology: a case study of the naturalized Asteraceae in Australia. NeoBiota 34: 1–20. https://doi.org/10.3897/neobiota.34.11139

Pagad S, Genovesi P, Carnevali L, Schigel D, McGeoch MA(2018) Introducing the Global Register of Introduced and Invasive Species. Scientific Data 5: 170202. https://doi.org/10.1038/sdata.2017.202

Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP (2010) Names are key to the big new biology. Trends Ecol Evol 25: 686–691. https://doi.org/10.1016/j.tree.2010.09.004

Pyšek P, Danihelka J, Sádlo J, Jr JC (2012) Catalogue of alien plants of the Czech Republic: checklist update, taxonomic diversity and invasion patterns. Preslia 84: 155–255.

Pyšek P, Pergl J, Essl F, Lenzner B, Dawson W, Kreft H, Weigelt P, Winter M, Kartesz J, Nishino M, Antonova LA, Barcelona JF, Cabezas FJ, Cárdenas D, Cárdenas-Toro J, Castańo N, Chacón E, Chatelain C, Dullinger S, Ebel AL, Figueiredo E, Fuentes N, Genovesi P, Groom QJ, Henderson L, Inderjit, Kupriyanov A, Masciadri S, Maurel N, Meerman J, Morozova O, Moser D, Nickrent D, Nowak PM, Pagad S, Patzelt A, Pelser PB, Seebens H, Shu W, Thomas J, Velayos M, Weber E, Wieringa JJ, Baptiste M, van Kleunen M (2017) Naturalized alien flora of the world: species diversity, taxonomic and phylogenetic patterns, geographic distribution and global hotspots of plant invasion. Preslia 89: 203–274. https://doi.org/10.23855/preslia.2017.203

Rees J, Cranston K (2017) Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodiversity Data Journal 5: e12581. https://doi.org/10.3897/BDJ.5.e12581

Richardson DM, Pyšek P, Rejmanek M, Barbour MG, Panetta FD, West CJ (2000) Naturalization and invasion of alien plants: concepts and definitions. Diversity and Distributions 6: 93–107. https://doi.org/10.1046/j.1472-4642.2000.00083.x

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, Blackburn TM, Dyer EE, Genovesi P, Hulme PE, Jeschke JM, Pagad S, Pyšek P, van Kleunen M, Winter M, Ansong M, Arianoutsou M, Bacher S, Blasius B, Brockerhoff EG, Brundu G, Capinha C, Causton CE, Celesti-Grapow L, Dawson W, Dullinger S, Economo EP, Fuentes N, Guénard B, Jäger H, Kartesz J, Kenis M, Kühn I, Lenzner B, Liebhold AM, Mosena A, Moser D, Nentwig W, 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, Walker K, Ward DF, Yamanaka T, Essl F (2018) Global rise in emerging alien species results from increased accessibility of new source pools. Proceedings of the National Academy of Sciences 115: E2264–E2273. https://doi.org/10.1073/pnas.1719429115

van der Aalst W, van Hee KM (2002) Workflow Management: Models, Methods, and Systems. MIT Press Cambridge, London. https://doi.org/10.7551/mitpress/7301.001.0001

van Kleunen M, Dawson W, Essl F, Pergl J, Winter M, Weber E, Kreft H, Weigelt P, Kartesz J, Nishino M, Antonova LA, Barcelona JF, Cabezas FJ, Cárdenas D, Cárdenas-Toro J, Castaño N, Chacón E, Chatelain C, Ebel AL, Figueiredo E, Fuentes N, Groom QJ, Henderson L, Inderjit, Kupriyanov A, Masciadri S, Meerman J, Morozova O, Moser D, Nickrent DL, Patzelt A, Pelser PB, Baptiste MP, Poopath M, Schulze M, Seebens H, Shu W-S, Thomas J, Velayos M, Wieringa JJ, Pyšek P (2015) Global exchange and accumulation of non-native plants. Nature 525: 100–103. https://doi.org/10.1038/nature14910

van Kleunen M, Pyšek P, Dawson W, et al. (2019) The Global Naturalized Alien Flora (GloNAF) database. Ecology 100: e02542.

Vanderhoeven S, Adriaens T, Desmet P, Strubbe D, Backeljau T, Barbier Y, Brosens D, Cigar J, Coupremanne M, De Troch R, Eggermont H, Heughebaert A, Hostens K, Huybrechts P, Jacquemart A-L, Lens L, Monty A, Paquet J-Y, Prévot C, Robertson T, Termonia P, Van De Kerchove R, Van Hoey G, Van Schaeybroeck B, 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

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, Hoen TAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18

Wilson JRU, Faulkner KT, Rahlao SJ, Richardson DM, Zengeya TA, van Wilgen BW (2018) Indicators for monitoring biological invasions at a national level. Journal of Applied Ecoloy 55: 2612–2620. https://doi.org/10.1111/1365-2664.13251

Supplementary material 1

Technical description and manual of the SInAS workflow implementation in R

Authors: Hanno Seebens, David A. Clarke, Quentin Groom, John R. U. Wilson, Emili García-Berthou, Ingolf Kühn, Mariona Roigé, Shyama Pagad, Franz Essl, Joana Vicente, Marten Winter, Melodie McGeoch

Data type: text

Explanation note: This document contains a detailed description of the implementation of the SInAS workflow in R and its application.

Link: https://doi.org/10.3897/neobiota.59.53578.suppl1

Supplementary material 2

Supplementary Tables S1–S4

Authors: Hanno Seebens, David A. Clarke, Quentin Groom, John R. U. Wilson, Emili García-Berthou, Ingolf Kühn, Mariona Roigé, Shyama Pagad, Franz Essl, Joana Vicente, Marten Winter, Melodie McGeoch

Data type: tables

Explanation note: Tables S1–S4 provide descriptions of variables used in the SInAS workflow and how these were mapped on the databases of the case study.

Link: https://doi.org/10.3897/neobiota.59.53578.suppl2

Supplementary material 3

Final output files of the case study applying the SInAS workflow

Authors: Hanno Seebens, David A. Clarke, Quentin Groom, John R. U. Wilson, Emili García-Berthou, Ingolf Kühn, Mariona Roigé, Shyama Pagad, Franz Essl, Joana Vicente, Marten Winter, Melodie McGeoch

Data type: tables

Explanation note: The zip contains the final output files of the application of the SInAS workflow in the case study. It includes the merged database, a full list of taxon names and the translated location names and event dates (first records).

Link: https://doi.org/10.3897/neobiota.59.53578.suppl3

Supplementary material 4

Intermediate output files of the case study applying the SInAS workflow

Authors: Hanno Seebens, David A. Clarke, Quentin Groom, John R. U. Wilson, Emili García-Berthou, Ingolf Kühn, Mariona Roigé, Shyama Pagad, Franz Essl, Joana Vicente, Marten Winter, Melodie McGeoch

Data type: tables

Explanation note: The zip file contains all intermediate output files, which represent the output of each individual step of the SInAS workflow applied in the case study.

Link: https://doi.org/10.3897/neobiota.59.53578.suppl4

Supplementary material 5

Unresolved entries of the case study applying the SInAS workflow

Authors: Hanno Seebens, David A. Clarke, Quentin Groom, John R. U. Wilson, Emili García-Berthou, Ingolf Kühn, Mariona Roigé, Shyama Pagad, Franz Essl, Joana Vicente, Marten Winter, Melodie McGeoch

Data type: tables

Explanation note: The zip file contains all files with unresolved records such as unmatched taxon names, missing location names, unresolved event dates or missing terms from the application of the workflow in the case study. These files could be used for cross-checking and correction errors and mis-matches to improve the final output.

Link: https://doi.org/10.3897/neobiota.59.53578.suppl5