NeoBiota 74: 75–91 (2022)
doi: 10.3897/neobiota.74.81082
https://neobiota.pensoft.net

SOFTWARE DESCRIPTION

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

DASCO: A workflow to downscale alien species checklists using occurrence records and to re-allocate species distributions across realms

Hanno Seebens[1], Ekin Kaplan[1,2,3]

1 Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25, 60325 Frankfurt, Germany2 BioInvasions, Global Change, Macroecology-Group, Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, Vienna 1030, Austria 3 Middle East Technical University, Department of Biology, 06800 Ankara, Turkey

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

Academic editor: Joana Vicente | Received 24 January 2022 | Accepted 27 April 2022 | Published 17 June 2022

Citation: 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. https://doi.org/10.3897/neobiota.74.81082

Abstract

Information about occurrences of alien species is often provided in so-called checklists, which represents lists of reported alien species in a region. In many cases, available checklists cover whole countries, which is too coarse for many analyses and limits capabilities of assessing status and trends of biological invasions. Information about point-wise occurrences is available in large quantities at online facilities such as GBIF and OBIS, which, however, do not provide information about the invasion status of individual populations. To close this gap, we here provide a semi-automated workflow called DASCO to downscale regional checklists using occurrence records obtained from GBIF and OBIS. Within the workflow, coordinate-based occurrence records for species listed in the provided regional checklists are obtained from GBIF and OBIS, and the status of being an alien population is assigned using the information in the provided checklists. In this way, information in checklists is made available at the local scale, which can then be re-allocated to any other spatial categorisation as provided by the user. In addition, habitats of species are determined to distinguish between marine, brackish, terrestrial, and freshwater species, which allows splitting the provided checklists to the respective realms and ecoregions. By using checklists of global databases, we showcase the usage of the DASCO workflow and revealed > 35 million occurrence records of alien populations in terrestrial and marine regions worldwide, which were back-transformed to terrestrial and marine regions for comparison. DASCO has the potential to be used as a basis for the widely applied species distribution models or assessments of status and trends of biological invasions at large geographic scales. The workflow is implemented in R and in full compliance with the FAIR data principles of open science.

Keywords
biological invasion, checklists, coordinates, distribution, downscaling, GBIF, marine ecoregions, neobiota, open science, workflow

Introduction

The amount of biodiversity data is increasing at an unprecedented pace (La Salle et al. 2016), with occurrence records provided by the Global Biodiversity Information Facility (GBIF) amounting to more than 2 billion records at the date of publication. Other online platforms such as the Ocean Biodiversity Information System (OBIS) are expanding likewise, although at lower levels. These platforms provide the by far largest collections of species occurrence records, which make them most useful for analysing the status and trends of biodiversity in general. The data on these platforms provided a basis for numerous analyses and biodiversity assessments but also exhibited distinct biases, gaps, and heterogeneity in quality and, therefore, should be handled with care to deal with these issues (Meyer et al. 2015; Hughes et al. 2021). Many of the recorded occurrences represent records of species outside their native range, so-called alien populations. However, these databases lack information about the status of invasion, which limits the capabilities to use the data for assessing trends in biological invasions.

As the number of biodiversity records increased, so did the number of records of alien populations collected in regional to global databases. Since 2015, at least seven new global databases of alien species records have been published: five of certain taxonomic groups such as alien plants (van Kleunen et al. 2019), birds (Dyer et al. 2017), mammals (Biancolini et al. 2021), amphibians and reptiles (Capinha et al. 2017) and macrofungi (Monteiro et al. 2020), and two major cross-taxonomic databases, one database on invasive alien species (Pagad et al. 2018) and one on years of first alien species' record (Seebens et al. 2017). Numerous collections at regional levels are available in addition. The standard format of alien species records is a checklist, which represents a list of species reported in a certain region, usually a country (Pyšek et al. 2012; Brundu and Camarda 2013). While these checklists provide a first overview of the distribution of alien species at larger geographic scales, the resolution is often too coarse to perform detailed analyses. For instance, the majority of alien species are still spreading despite their first introduction being decades or centuries ago (Seebens et al. 2021), but the availability of distribution records only at a regional scale distinctly hampers the assessment of the dynamics of spread and severely limits the possibility to predict the future spread and hot spots of alien species occurrences.

The rise of biodiversity data poses new challenges to researchers as the processing of data becomes increasingly complex and time-consuming. As the steps of data processing are often similar in different projects, researchers spent much time on developing very similar approaches multiple times, which is inefficient. In addition, the complexity of data processing requires making many minor decisions of how to handle and modify data, which are usually not reported in the method section of a scientific publication. As a consequence, studies and assessments are non-transparent and not reproducible, which reduces trust in scientific results (Franz and Sterner 2018). It is therefore of rising importance to publish all steps of data processing, the so-called workflows (Hardisty and Roberts 2013). With the rise in data volumes and complexities of data processing, it also becomes crucial to make workflows accessible to others (Guralnick et al. 2007), which provides the opportunity to document all steps of the process accurately, to make studies transparent and reproducible, to increase efficiency in science by allowing others to use the workflow, and to ultimately increase trust in study results.

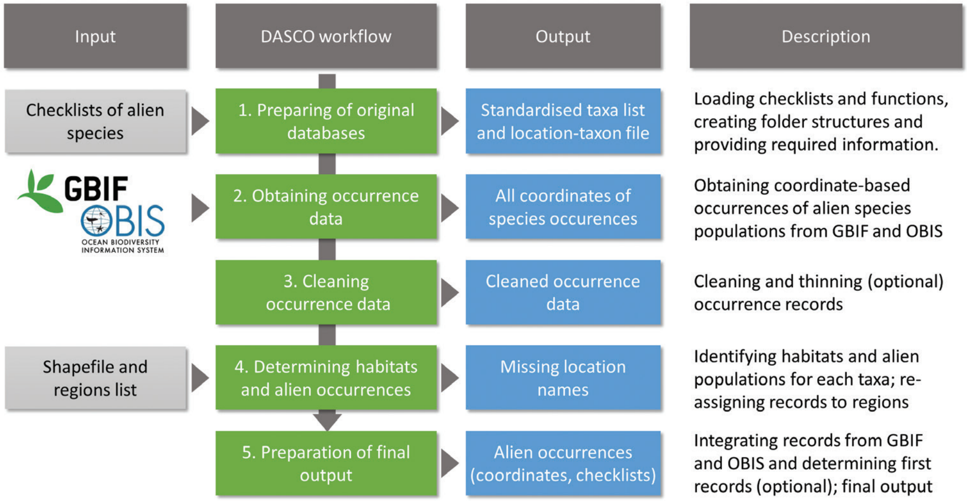
In recent years, much progress has been made on developing standards, workflows, and infrastructures for biodiversity information. For example, a standard terminology for biodiversity information called Darwin Core (https://dwc.tdwg.org/) has been developed, which allows sharing data more easily (Groom et al. 2019). Workflows (i.e., technical pipelines to process data) have been proposed and developed to clean biodiversity data (Zizka et al. 2019) and to transform the massive amount of occurrence data into workable formats (Guralnick et al. 2007; Jetz et al. 2019). Standard measures of biodiversity have been proposed and accepted, such as the Essential Biodiversity Variables (Pereira et al. 2013) and a range of indicators to actually measure biodiversity change. However, most of these advancements relate to biodiversity information in general, while the specifics of biological invasions were often not taken into account, and similar developments in invasion ecology are lagging behind the general trends. Efforts have been made in some parts. For example, the Darwin Core terminology has been extended to capture aspects of the status of biological invasions (Groom et al. 2019), workflows have been published to integrate global databases (Seebens et al. 2020), and indicators have been developed to measure and visualise trends in changes of biological invasions (Wilson et al. 2018), but still, information about the status of alien species population is usually provided on national scales with all the limitation inherent in such a coarse scale, although higher resolved data are available.

Here, we provide a workflow that integrates the strengths of both the comprehensiveness of point-wise occurrence records provided by GBIF and OBIS and information on invasion status provided in checklists. While GBIF provided the by far largest amount of occurrence data, OBIS represents a platform gathering information about mostly marine species occurrences. Their combination therefore provides a comprehensive compilation of species occurrences across realms. The ultimate goal of applying the workflow is to obtain occurrence records of alien populations with associated coordinates at large extent. By combining regional checklists and occurrence records, the information provided at coarse geographic scale such as regional checklists can be transferred to a finer geographic scale of local occurrences, a process often called 'downscaling' as used in e.g. climate science. Hence, the workflow can be used to downscale alien species checklists using occurrence records, and is therefore called 'DASCO', but also to re-allocate species occurrences to different delineations of regions or realms to generate checklists at alternative spatial resolutions. For instance, a single checklist may contain species from different realms, biomes, or ecotypes. By using coordinate-based occurrence records, it is then possible to split the checklists and assign species to, for example, bordering coastal areas or ecotypes such as mountainous areas within the respective region, and to generate checklists only for those areas with a resolution, which may differ from the original checklist.

In a case study, we showcase the application of the workflow at a global scale using the largest global database of alien species occurrences based on regional checklists. This case study provides an overview of the records of alien species populations globally distinguished between terrestrial, marine, and freshwater species. The DASCO workflow is fully implemented in the open-source language R (version 4.1.3, R Core Team 2022) and is published together with this article. The workflow was designed in a way that allows other users to modify and apply the scripts to their respective needs, for example, by providing their own region delineations for aggregating the occurrence data.

The DASCO workflow

The DASCO workflow is structured in a sequence of five steps of data processing (Fig. 1): 1) preparing of input data sets and folder structure, 2) obtaining occurrence records of species from GBIF and OBIS, 3) cleaning obtained occurrence records, 4) determining the invasion status (i.e., alien) of the populations, and finally 5) preparing the final output. The steps are executed in sequence and each produces output files, which are used as input of the next step. This enables the application of individual steps in isolation without the need to run the full workflow in all cases.



Input	DASCO workflow	Output	Description
Checklists of alien species	1. Preparing of original databases	Standardised taxa list and location-taxon file	Loading checklists and functions, creating folder structures and providing required information.
GBIF OBIS OCEAN BIODIVERSITY INFORMATION SYSTEM	2. Obtaining occurrence data	All coordinates of species occurences	Obtaining coordinate-based occurrences of alien species populations from GBIF and OBIS
	3. Cleaning occurrence data	Cleaned occurrence data	Cleaning and thinning (optional) occurrence records
Shapefile and regions list	4. Determining habitats and alien occurrences	Missing location names	Identifying habitats and alien populations for each taxa; re-assigning records to regions
	5. Preparation of final output	Alien occurrences (coordinates, checklists)	Integrating records from GBIF and OBIS and determining first records (optional); final output

Figure 1. Overview of the DASCO workflow. The workflow consists of five steps (green boxes), which are executed in sequence. It requires input from external sources (column 'Input') and exports a series of output files (blue boxes) to document the process, to provide intermediate output results, and the final output files.

The essential requirements for executing the workflow are the original database of alien taxa, which is organised as a checklist at any scale, a shapefile of the polygons of the regions, R installed on a computer, and a GBIF account. A detailed description of the workflow, requirements for running the workflow, and technical descriptions of the individual functions are available in the DASCO manual, which is available as an R Markdown file together with the code (https://doi.org/10.5281/zenodo.5841930) and as a pdf (see Suppl. material 1). An overview of the individual steps of the workflow is presented in the following:

Step 1: Preparation of database

In the first step of the DASCO workflow, checklists of alien species are imported and prepared for further processing. A checklist represents a list of species, which are known to occur in a certain region. Usually, regions (also called 'location') represent a country, an island, or a nature reserve, but it could be any area of any size. Column headers of the columns containing taxon names, locations, and first record are standardised according to Darwin Core terminology following Groom et al. (2019). In addition, location names are standardised according to an associated translation table. This translation table can be modified or replaced by the user to obtain a different set of location names. A standardised spreadsheet table of location-taxon records and a list of all taxa are exported. Note that taxon names are not standardised, as this could be done using other workflows (Seebens et al. 2020; Grenié et al. 2022), which could be applied before the application of DASCO.

Step 2: Obtaining occurrence data

In the second step of the DASCO workflow, available occurrence records for each species, which are listed in the checklists provided in step 1, are obtained from GBIF and OBIS. All available occurrence records are downloaded irrespective of their location or invasion status of the respective population. Depending on the length of the species list, this may result in large amounts of data, particularly for GBIF data, which may be difficult to process in one step. Thus, the number of available records on GBIF for each species is determined beforehand. By default, the request to GBIF is automatically split into three chunks, which can be processed in parallel using a single GBIF account. If the total number of records is large, the user can provide multiple accounts, the taxa are split accordingly, and individual requests for download are sent for each chunk to obtain data sets of manageable sizes. This step requires one or multiple accounts on GBIF to allow processing multiple chunks of data simultaneously (see the DASCO manual for further details).

Once the GBIF files are ready for download, they will be downloaded to a local folder. GBIF provides digital unique identifiers (DOI) for each query, which are exported by the workflow and should be kept and provided to ensure transparency and reproducibility. The downloaded files are decompressed, and an initial cleaning is conducted by removing duplicated, empty and non-numeric entries of the columns ‘speciesKey’, ‘decimalLatitude,’ and ‘decimalLongitude.’ In addition, obviously wrong coordinates with values being outside the coordinate systems are removed (original records are kept for cross checking). Finally, all records indicated as ‘FOSSIL_SPECIMEN’ are removed.

For OBIS, the number of available occurrence records is usually much lower compared to GBIF. Therefore, it is not necessary to perform initial checks and to split download requests. Thus, all available records for species of the provided checklists are directly imported into R. Duplicated records and records, which are indicated as ‘FossilSpecimen’, are removed. OBIS does not provide a DOI for individual queries. Lists of all records from GBIF and OBIS are exported and saved locally.

Step 3: Cleaning occurrence data

The third step represents the most computer- and time-intensive part of the workflow as it contains the cleaning of the obtained occurrence records. Occurrence records provided on GBIF and OBIS are prone to errors and uncertainties due to inaccurate measurements or wrong entries and therefore require cleaning. First, inaccurate coordinates with fewer than two digits after the comma are removed. This is considered to be a minimum requirement, and a higher resolution might be desired depending on the geographic resolution of the study, while for large-scale databases, such accuracy should be sufficient. Subsequently, seven tests of validation are applied to identify wrong coordinates. The tests are provided by the R package ‘CoordinateCleaner,’ which was specifically designed to validate occurrence records provided by platforms such as GBIF (Zizka et al. 2019). These tests involve checking whether, for example, coordinates represent centroids or capitals of countries, the location of large biodiversity institutions or the headquarter of GBIF rather than actual species populations. The most important test for our purpose represents the check for outliers, which identifies records that are located at large distances to the majority of records. These records might be a result of misspecifications or erroneous entries. Records flagged as potentially wrong entries by the tests are removed from the list, which - based on experiences - represents around 5% of records. This resulted in a more conservative estimate of the actual species occurrence. These tests are applied to records of both platforms. The user has the opportunity to check the removal of records by comparing the original downloaded occurrence files with the output file of the workflow.

Due to the sheer amount of data provided by GBIF, conducting the outlier test could be time- and memory-consuming. Many of the records represent multiple counts of the same species within a narrow geographic range, which would not add new information to our workflow. To improve the efficiency and speed of the workflow, we allowed for the thinning of records to reduce the workload. Thinning was done by rounding the coordinates to the second digit after the comma, keeping only one record (but the original, not rounded coordinates) for this occurrence, and removing others. Depending on the focus of the study, thinning could be done to finer geographic scales or disabled at all. Thinning is disabled by default for records provided by OBIS but can be turned on if required.

Step 4: Determining alien occurrences and habitats

Within the fourth step of the DASCO workflow, the cleaned occurrence records and the original checklists are used to identify alien populations. This requires having a shapefile with the same region borders as provided in the checklists. Only occurrence records were kept, which were located in the regions, where the respective species was classified as being alien. In this way, it is ensured that the information about the invasion status of being an alien taxon in a certain location has been assigned to the occurrence records. Records falling outside those regions were removed. As a default, a shapefile of country borders, large islands, and marine ecoregions is provided and used. Only those combinations of a taxon and a region are kept in the workflow if at least three occurrence records within the respective region are available for the taxon. Fewer numbers of records per taxon-region combination are considered to be too uncertain and removed. The emergence of region names of the checklists, which are not matching the names provided in the shapefile, will produce a warning and an export of mismatching region names.

Checklists often contain taxa of different habitats (e.g., terrestrial, marine, freshwater). As the region of record provided in the shapefile is often a terrestrial region, such as the land of a country or island, occurrences of recorded marine taxa often fall outside the provided polygons. The availability of coordinate-based occurrence records now provides the opportunity to specify the coastal area of the region, where the taxon actually occurs. In addition to occurrence records, this requires the determination of habitats for each taxon, a delineation of marine coastal regions, and knowledge about borders of land and marine coastal regions. We, therefore, provide a list of regions and their bordering marine ecoregions based on the classification provided by Spalding et al. (2007). Occurrence records of taxa, which have been identified as being marine and alien on a regional checklist, are considered to describe alien populations in the neighbouring marine ecoregions. Thus, occurrences of a marine taxon are assigned to a marine ecoregion only if the taxon is listed as being alien for the region (i.e., a country) and has at least three occurrence records in the respective marine ecoregion.

As records of many taxa, which are actually not marine, fall into polygons of marine ecoregions, an additional step of determining habitats of a taxon has been included. For each taxon, information about the habitat is obtained from the online databases WoRMS (WoRMS Editorial Board 2022), FishBase (Froese and Pauly 2021), and Sea LifeBase (www.sealifebase.ca) if entries for the taxon exist. Multiple entries are allowed for species capable of moving between habitats. Only records of taxa identified as being marine are assigned to a marine ecoregion. As habitat information provided a number of false entries, the following taxon groups were excluded from marine ecoregions: Vascular plants, insects, spiders, bryophytes, birds, amphibians, and mammals. In addition, only those species were considered as being marine, which were explicitly mentioned as such in the aforementioned databases or in the databases provided by the user. Marine mammals are excluded because, up to now, no introduction of a marine mammal has been reported. These restrictions may result in the removal of actual true records, but overall will ensure avoiding large numbers of false entries in the final output, which is preferred. Other habitat types were taken as provided by the online databases or the input checklist without any test using occurrence records, because occurrence data often do not provide the accuracy to distinguish between, for example, terrestrial and freshwater habitats. Habitat information can also be provided as a separate column in the input data set.

Two data sets are exported from step 4: A list of occurrence records with coordinates for alien populations with the associated name of the region and a list of taxon-region combinations. The latter represents checklists as provided in the original input file, which is now cross-checked by records from GBIF and OBIS and may include new regions such as marine ecoregions. Providing different shapefiles would allow re-assigning the occurrences to an alternative set of regions.

Step 5: Merging data sets and finalising the output

In the last step of the workflow, data sets of occurrences of alien species at a regional scale will be merged and prepared for the final output. Steps 2–4 are split into parallel strands for GBIF and OBIS, which are merged here to obtain a single output. Duplicated records are removed. If information about the year of the first record has been provided, it will be assigned at this step to the respective taxon and region. If multiple first records exist due to, e.g., the usage of a different geographic classification, the earliest first record is selected.

A case study

We showcase the application of the DASCO workflow using the SInAS database. The SInAS database represents an output from another workflow (i.e., the SInAS workflow; Seebens et al. 2020) designed to integrate databases of alien species occurrences based on checklists in a semi-automated and transparent way of standardisation and integration. Here, we use version 2.4.1 of the SInAS database (https://doi.org/10.5281/zenodo.5562892), which results from the integration of seven global databases of alien species occurrences: Five taxonomic databases, namely for vascular plants (GloNAF; van Kleunen et al. 2019), birds (GAVIA; Dyer et al. 2017), mammals (Biancolini et al. 2021), macrofungi (Monteiro et al. 2020) and amphibians and reptiles (Capinha et al. 2017), and two cross-taxonomic databases being one about temporal information of first recording (FirstRecords; Seebens et al. 2017) and one about invasive alien species (GRIIS; Pagad et al. 2022). All seven databases are based on checklists of regional (mostly country) scale. By applying the SInAS workflow, the terminologies, taxonomies, regional delineations, and event dates of the individual databases were standardised and the standardised databases were merged into the SInAS database. This version of the SInAS database contains 175.980 records of 39.191 alien taxa occurring in 264 non-overlapping regions worldwide. As the SInAS database is organised as a collection of checklists for regions, it can be directly used as input for the DASCO workflow.

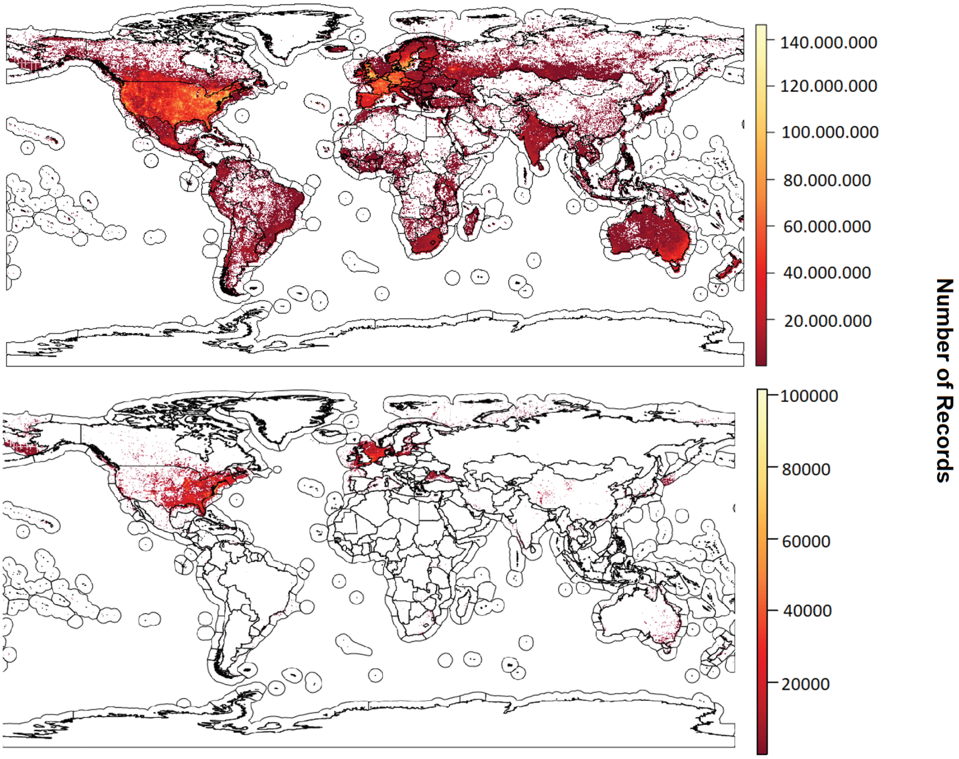


Figure 2. The number of records of alien populations obtained from GBIF (top) and OBIS (bottom).

Applying the DASCO workflow to the SInAS database required processing large amounts of occurrence data, which altogether took around four days, with the longest step being the cleaning of the GBIF data. The application of the DASCO workflow resulted in a total of 35.666.064 cleaned coordinate-based occurrence records of alien populations of 17,424 taxa (Fig. 2). The vast majority of records (99%) was obtained from GBIF, and only a comparatively small fraction stemmed from OBIS. Records of both databases are heavily biased towards Europe, North America, and Australia.

While checklists often provide comprehensive lists of taxa, more detailed information about the exact occurrences of populations is limited to a distinctly lower number of taxa. Consequently, while applying the workflow, the number of taxon-region combinations likely reduces due to the lower number of taxa in GBIF and OBIS and information gaps. Indeed, information about the occurrence of alien populations was only available for 17,424 alien taxa, which is 44% of the number of species as provided in the original database.

The application of the DASCO workflow may introduce new or intensify already existing geographic and taxonomic biases due to biases of data provided by the online platforms. Although the application of the workflow resulted in a drop in available records, the proportions of reduction are fairly constant across all large-scale regions

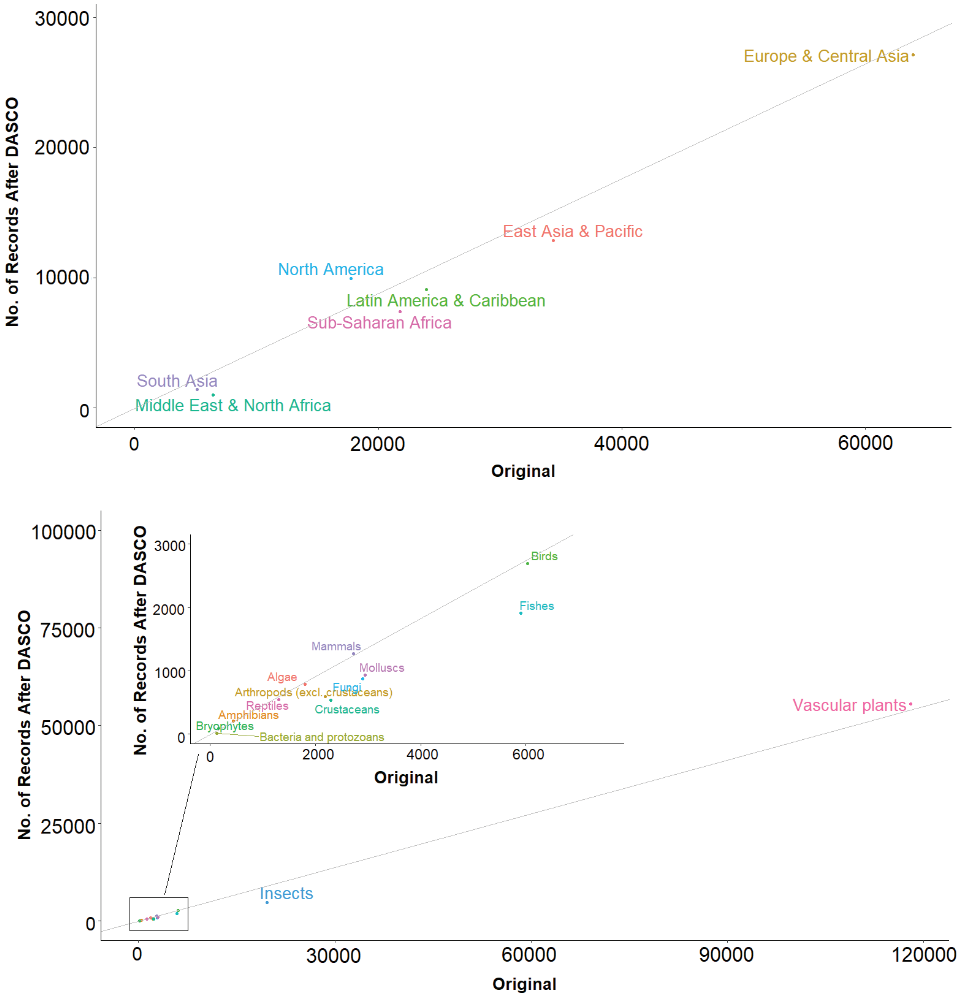


Figure 3. The number of taxon-region combinations before (x-axes, 'Original') and after (y-axes, 'DASCO') applying the DASCO workflow for different regions (upper panel) taxonomic groups (lower panel).

with an average decline of 64% (Fig. 3), with the highest and lowest values reported for the Middle East & North Africa (84.1%) and Europe & Central Asia (57.6%), respectively. Overall, there is no indication that the application of the workflow increased the geographic bias, which is certainly inherent in the original databases. Comparing records of taxonomic groups revealed a stronger decline for insects, fishes, molluscs, crustaceans, and fungi, while the decline was lower for vascular plants.

Habitat information was obtained for 21.605 taxa (64% of the requested number of 33.587 taxa). The majority of habitat records were terrestrial (58%), followed by marine (13%), freshwater (9%), and brackish (2%) (Fig. 4). Years of first records were available for 42% of all taxon-region combination. Long-term trends of the number of new alien taxa per five years revealed a clear increasing trend of the rate of first records

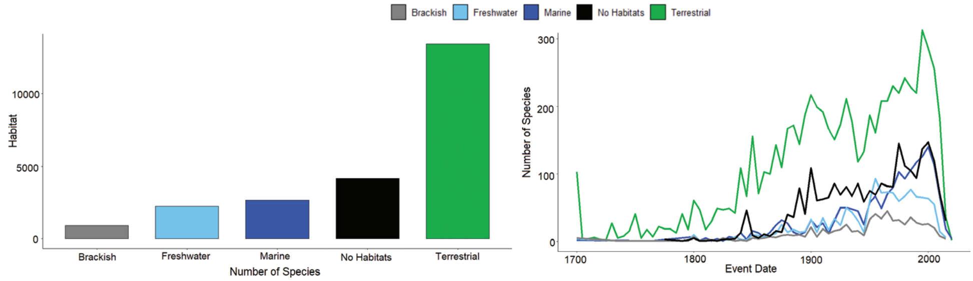


Figure 4. Overview of obtained habitat information. Shown are the total number of taxa with obtained habitat information (left panel) and long-term trends of alien taxon numbers distinguished by habitats (right panel).

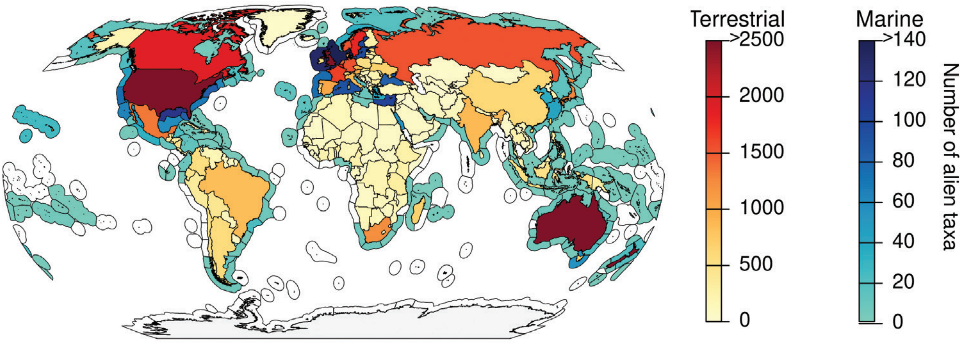


Figure 5. Map of the number of recorded alien taxa for terrestrial (freshwater + terrestrial) and marine (marine + brackish) taxa as obtained by the DASCO workflow.

until 2005, particularly for terrestrial and marine alien taxa, while rates for freshwater and brackish taxa saturated after ca. 1950 and slightly declined until today (Fig. 4).

The application of the DASCO workflow allowed the separation of checklists by habitats and the representation of alien taxon numbers for terrestrial regions (i.e., terrestrial + freshwater) and coastal marine regions (marine + brackish) (Fig. 5). For both terrestrial and marine regions, a geographic bias towards Europe, North America, and Australasia becomes apparent. The low numbers of available records, particularly for Africa, Central Asia, and many marine ecoregions, makes it difficult to identify any variation across regions and likely results from the lack of records in the used data sources.

Discussion

Checklists of alien taxa provide valuable and often comprehensive information about the invasion status of populations at regional levels, while online portals such as GBIF and OBIS provide tremendous amounts of data at higher spatial resolution. Here, we provide a workflow to integrate the advantages of both sources by assigning the invasion status obtained from checklists to occurrence records obtained from online portals. The DASCO workflow allows downscaling regional checklists to coordinate-based occurrences, which can then be used to re-assign occurrences to any categorisation provided by the user. In this way, the information provided in checklists, which are bound to a fixed delineation, is made accessible for a range of different purposes, including the assessment of biological invasions at resolutions, deviating from the original checklists. By applying the DASCO workflow, downscaling and re-assignment is done in a standardised, reproducible, and transparent way and in full compliance with the FAIR data principles (Wilkinson et al. 2016).

Our case study of applying the DASCO workflow to the SInAS database of alien taxa checklists resulted in a comprehensive compilation of coordinate-based occurrence records of alien populations. However, the distribution of records is highly biased towards a few well-sampled regions such as Europe, North America, Australia, and New Zealand, while particularly countries in Africa except South Africa, and Central Asia are highly under-represented (Fig. 2). This bias is even more pronounced for records obtained from OBIS. Aggregating the records back to the original regional delineation revealed a global pattern of alien taxa occurrences, which is very similar to what has been published elsewhere (Dyer et al. 2017; Pyšek et al. 2017). This is not surprising as both representations are based on the same data, but show that the application of the DASCO workflow does not distort the original maps except that the total numbers of taxa are lower.

For marine ecoregions, comparable global maps of alien marine taxa do not exist. Bailey et al. (2020) published the most recent and comprehensive compilation of marine alien taxa, which, however, still covers only approximately half the world's ecoregion at a coarser resolution than provided here. But the overall patterns are similar to our results, although distinctly higher numbers of marine alien taxa can be expected for most marine ecoregions except probably for European and North American coastal waters. Our case study highlights that downscaling and re-allocating alien species occurrences using the DASCO workflow could provide a promising way to form a basis for large-scale assessments of biological invasions for regions, which are not yet well covered in global analyses.

The DASCO workflow is limited in different ways, which should be taken into account. First of all, the output of the workflow highly depends on the information provided in online sources. As this information is often geographically and taxonomically biased (Fig. 2; Meyer et al. 2016; Rocha-Ortega et al. 2021), obtained records are likely biased as well, which, however, depends on the taxon and region considered. While for well-sampled regions and taxa, a reduction might be low, the loss of information might be very high for under-sampled cases such as microorganisms or Central Africa. In addition, provided records might be of low quality, including false or imprecise coordinates (Jin and Yang 2020), and thus obtained records should be handled with care (Zizka et al. 2019). This is particularly problematic at small geographic scales, where imprecise coordinates can make a big difference when, for example, it is unclear whether a taxon is found inside or outside a nature reserve. We included a number of tests to identify imprecise and wrong entries, but these likely do not remove all faulty records. These errors became less influential at larger scales, and thus results from the application of the DASCO workflow should be treated more carefully with increasing spatial resolution of the analysis. Furthermore, as there is no single comprehensive source of habitat information for taxa, the habitat type could not be identified for many taxa, particularly aquatic ones. All of these limitations can only be solved by increasing the amount of information provided by online sources, which is an ongoing but long-lasting process. Additional software packages and workflows have been developed to identify and, to some degree, correct errors in spatial information (Mathew et al. 2014; Jin and Yang 2020), which could be applied in addition.

Another limitation of the workflow is that it currently cannot discriminate native from alien populations. Although the workflow can identify alien populations based on regional checklists, this does not automatically mean that all records not classified as being alien belong to native populations. It might be that some records refer to alien populations, which are not included in the regional checklists. It therefore remains unsafe to classify native populations using our workflow. Still, this can cause an increase in false positive records for species, which have both native and alien ranges within the same region. Such species might be considered as being alien in the regional checklist. In this case, the workflow would assign all records within the region the status of being alien, although some populations may in fact be native. This depends on the scale, at which the checklists are provided, and can only be avoided by using checklists at subnational scale for large countries to distinguish e.g. federal states and islands.

The DASCO workflow has been designed in the context of biological invasions, but its use is not limited to this area, as coordinate-based occurrences of any kind of taxon checklist can be downscaled and re-allocated across varying delineations and realms. In addition, parts of the workflow could be applied in isolation. For example, obtaining and cleaning large amounts of GBIF records in a convenient and transparent way is likely of interest for many users for various purposes. As other potential applications, obtained records of alien taxa could be used to identify native populations, and the integration of habitat information could potentially be of interest for other research studies.

By using available and open workflows, such work becomes more efficient because work does not have to be repeated as it is often done right now in parallel projects. With the increase in the amount of data, developing and sharing workflows such as DASCO becomes more and more important to make unstructured data accessible in a reproducible and transparent way, which ultimately will increase trust in scientific outcomes (Franz and Sterner 2018).

Data and code availability

All necessary files for running the DASCO workflow, such as R scripts, the shapefile, and the marine-terrestrial region file, are available for public use at Github with version control (https://github.com/hseebens/DASCOworkflow) and releases are stored on Zenodo (https://doi.org/10.5281/zenodo.5841930). The SInAS database, which represents the input data set for the case study, is available online (https://doi.org/10.5281/zenodo.5562892). The occurrence records, which are exported by the DASCO workflow for the case study, are provided online together with a list of identifiers of original GBIF downloads (https://doi.org/10.5281/zenodo.6458083).

Acknowledgements

The research was funded through the 2017–2018 Belmont Forum and BiodivERsA joint call for research proposals, under the BiodivScen ERA-Net COFUND programme, and with the funding organisation BMBF (grant number 16LC1807A).

References

Bailey SA, Brown L, Campbell ML, Canning-Clode J, Carlton JT, Castro N, Chainho P, Chan FT, Creed JC, Curd A, Darling J, Fofonoff P, Galil BS, Hewitt CL, Inglis GJ, Keith I, Mandrak NE, Marchini A, McKenzie CH, Occhipinti-Ambrogi A, Ojaveer H, Pires-Teixeira LM, Robinson TB, Ruiz GM, Seaward K, Schwindt E, Son MO, Therriault TW, Zhan A (2020) Trends in the detection of aquatic non-indigenous species across global marine, estuarine and freshwater ecosystems: A 50-year perspective. Diversity & Distributions 26(12): 1780–1797. https://doi.org/10.1111/ddi.13167

Biancolini D, Vascellari V, Melone B, Blackburn TM, Cassey P, Scrivens SL, Rondinini C (2021) DAMA: The global distribution of alien mammals database. Ecology 102(11): e03474. https://doi.org/10.1002/ecy.3474

Brundu G, Camarda I (2013) The Flora of Chad: A checklist and brief analysis. PhytoKeys 23(0): 1–18. https://doi.org/10.3897/phytokeys.23.4752

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 & Distributions 23(11): 1313–1322. https://doi.org/10.1111/ddi.12617

Dyer EE, Redding DW, Blackburn TM (2017) The global avian invasions atlas, a database of alien bird distributions worldwide. Scientific Data 4(1): e170041. https://doi.org/10.1038/sdata.2017.41

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

Froese R, Pauly D [Eds] (2021) FishBase. World Wide Web electronic publication. www.fishbase.org

Grenié M, Berti E, Carvajal-Quintero J, Dädlow GML, Sagouis A, Winter M (2022) Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. Methods in Ecology and Evolution 2022: 1–14. https://doi.org/10.1111/2041-210X.13802

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

Guralnick RP, Hill AW, Lane M (2007) Towards a collaborative, global infrastructure for biodiversity assessment. Ecology Letters 10(8): 663–672. https://doi.org/10.1111/j.1461-0248.2007.01063.x

Hardisty A, Roberts D (2013) A decadal view of biodiversity informatics: Challenges and priorities. BMC Ecology 13(1): e16. https://doi.org/10.1186/1472-6785-13-16

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(9): 1259–1269. https://doi.org/10.1111/ecog.05926

Jetz W, McGeoch MA, Guralnick R, Ferrier S, Beck J, Costello MJ, Fernandez M, Geller GN, 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 & Evolution 3(4): 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

La Salle J, Williams KJ, Moritz C (2016) Biodiversity analysis in the digital era. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 371(1702): e20150337. https://doi.org/10.1098/rstb.2015.0337

Mathew C, Güntsch A, Obst M, Vicario S, Haines R, Williams A, 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

Meyer C, Kreft H, Guralnick R, Jetz W (2015) Global priorities for an effective information basis of biodiversity distributions. Nature Communications 6(1): e8221. https://doi.org/10.1038/ncomms9221

Meyer C, Weigelt P, Kreft H (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecology Letters 19: 992–1006. https://doi.org/10.1111/ele.12624

Monteiro M, Reino L, Schertler A, Essl F, Figueira R, Ferreira M, Capinha C (2020) A database of the global distribution of alien macrofungi. Biodiversity Data Journal 8: e51459. https://doi.org/10.3897/BDJ.8.e51459

Pagad S, Genovesi P, Carnevali L, Schigel D, McGeoch MA (2018) Introducing the Global Register of Introduced and Invasive Species. Scientific Data 5(1): e170202. https://doi.org/10.1038/sdata.2017.202

Pagad S, Bisset S, Genovesi P, Groom Q, Hirsch T, Jetz W, Ranipeta A, Schigel D, Sica YV, McGeoch MA (2022) The Global Register of Introduced and Invasive Species: Country Compendium. bioRxiv: 2022.04.19.488841. https://doi.org/10.1101/2022.04.19.488841

Pereira HM, Ferrier S, Walters M, Geller GN, Jongman RHG, Scholes RJ, Bruford MW, Brummitt N, Butchart SHM, Cardoso AC, Coops NC, Dulloo E, Faith DP, Freyhof J, Gregory RD, Heip C, Hoft R, Hurtt G, Jetz W, Karp DS, McGeoch MA, Obura D, Onoda Y, Pettorelli N, Reyers B, Sayre R, Scharlemann JPW, Stuart SN, Turak E, Walpole M, Wegmann M (2013) Essential Biodiversity Variables. Science 339(6117): 277–278. https://doi.org/10.1126/science.1229931

Pyšek P, Chytrý M, Pergl J, Sádlo J, Wild J (2012) Plant invasions in the Czech Republic: current state, introduction dynamics, invasive species and invaded habitats. Preslia 84: 575–629. http://www.muni.cz/research/publications/993057

Pyšek P, Pergl J, Essl F, Lenzner B, Dawson W, Kreft H, Weigelt P, Winter M, Kartesz J, Nishino M, Antonova LA, Barcelona JF, Cabesaz FJ, Cárdenas D, Cárdenas-Toro J, Castaño 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 MP, 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(3): 203–274. https://doi.org/10.23855/preslia.2017.203

R Core Team (2022) R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.r-project.org/

Rocha-Ortega M, Rodriguez P, Córdoba-Aguilar A (2021) Geographical, temporal and taxonomic biases in insect GBIF data on biodiversity and extinction. Ecological Entomology 46(4): 718–728. https://doi.org/10.1111/een.13027

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, Stajerová 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(1): e14435. 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, Blackburn TM, Hulme PE, Kleunen M, Liebhold AM, Orlova-Bienkowskaja M, Pyšek P, Schindler S, Essl F (2021) Around the world in 500 years: Inter-regional spread of alien species over recent centuries. Global Ecology and Biogeography 30(8): 1621–1632. https://doi.org/10.1111/geb.13325

Spalding MD, Fox HE, Allen GR, Davidson N, Ferdaña ZA, Finlayson MAX, Halpern BS, Jorge MA, Lombana AL, Lourie SA, Martin KD, Manus MC, Molnar J, Recchia CA, Robertson J (2007) Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. Bioscience 57(7): 573–583. https://doi.org/10.1641/B570707

van Kleunen M, Pyšek P, Dawson W, Essl F, Kreft H, Pergl J, Weigelt P, Stein A, Dullinger S, König C, Lenzner B, Maurel N, Moser D, Seebens H, Kartesz J, Nishino M, Aleksanyan A, Ansong M, Antonova LA, Barcelona JF, Breckle SW, Brundu G, Cabezas FJ, Cárdenas D, Cárdenas-Toro J, Castaño N, Chacón E, Chatelain C, Conn B, Sá Dechoum M, Dufour-Dror J, Ebel AL, Figueiredo E, Fragman-Sapir O, Fuentes N, Groom QJ, Henderson L, Inderjit, Jogan N, Krestov P, Kupriyanov A, Masciadri S, Meerman J, Morozova O, Nickrent D, Nowak A, Patzelt A, Pelser PB, Shu W, Thomas J, Uludag A, Velayos M, Verkhosina A, Villaseñor JL, Weber E, Wieringa JJ, Yazlık A, Zeddam A, Zykova E, Winter M (2019) The Global Naturalized Alien Flora (GloNAF) database. Ecology 100: e02542. https://doi.org/10.1002/ecy.2542

Wilkinson MD, Dumontier M, Aalbersberg Ij J, 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, 't Hoen PA, 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(1): e160018. https://doi.org/10.1038/sdata.2016.18

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

WoRMS Editorial Board (2022) World Register of Marine Species. https://doi.org/10.14284/170

Zizka A, Silvestro D, Andermann T, Azevedo J, Duarte Ritter C, Edler D, Farooq H, Herdean A, Ariza M, Scharn R, Svantesson S, Wengström N, Zizka V, Antonelli A (2019) CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. Methods in Ecology and Evolution 10(5): 744–751. https://doi.org/10.1111/2041-210X.13152

Supplementary material 1

Manual of DASCO

Authors: Hanno Seebens, Ekin Kaplan

Data type: PDF file

Explanation note: Manual of DASCO: A workflow to down-scale alien species checklists using occurrence records and to re-allocate species distributions across realms.

Link: https://doi.org/10.3897/neobiota.74.81082.suppl1