Received: 24 May 2023 | Accepted: 14 January 2024

DOI: 10.1002/aps3.11575

Applications in Plant Sciences

SOFTWARE NOTE

Geographic And Taxonomic Occurrence R-based Scrubbing (gatoRs): An R package and workflow for processing biodiversity data

Natalie N. Patten[1] | **Michelle L. Gaynor**[2,3] | **Douglas E. Soltis**[2,3] | **Pamela S. Soltis**[2]

[1]Department of Mathematics, University of Florida, Gainesville 32611, Florida, USA

[2]Florida Museum of Natural History, University of Florida, Gainesville 32611, Florida, USA

[3]Department of Biology, University of Florida, Gainesville 32611, Florida, USA

Correspondence

Natalie N. Patten, Department of Mathematics, The Ohio State University, Columbus, Ohio 43210, USA.
Email: natalienpatten@gmail.com

Present address

Natalie N. Patten, Department of Mathematics, The Ohio State University, Columbus, Ohio 43210, USA.

This article is part of the special issue "Resilient botany: Innovation in the face of limited mobility and resources."

Abstract

Premise: Digitized biodiversity data offer extensive information; however, obtaining and processing biodiversity data can be daunting. Complexities arise during data cleaning, such as identifying and removing problematic records. To address these issues, we created the R package Geographic And Taxonomic Occurrence R-based Scrubbing (gatoRs).

Methods and Results: The gatoRs workflow includes functions that streamline downloading records from the Global Biodiversity Information Facility (GBIF) and Integrated Digitized Biocollections (iDigBio). We also created functions to clean downloaded specimen records. Unlike previous R packages, gatoRs accounts for differences in download structure between GBIF and iDigBio and allows for user control via interactive cleaning steps.

Conclusions: Our pipeline enables the scientific community to process biodiversity data efficiently and is accessible to the R coding novice. We anticipate that gatoRs will be useful for both established and beginning users. Furthermore, we expect our package will facilitate the introduction of biodiversity-related concepts into the classroom via the use of herbarium specimens.

KEYWORDS

basis cleaning, biodiversity data download, GBIF, herbaria, iDigBio, locality cleaning, spatial correction, taxonomic harmonization

The digitization of biodiversity data has greatly improved the accessibility of specimens stored in natural history collections, leading to novel research in many areas (e.g., Page et al., 2015; Soltis and Soltis, 2016; Soltis, 2017; Bakker et al., 2020). Data in aggregated biodiversity databases such as the Global Biodiversity Information Facility (GBIF: https://www.gbif.org/) and Integrated Digitized Biocollections (iDigBio: https://www.idigbio.org/) are standardized using a framework known as Darwin Core (DwC) (Wieczorek et al., 2012), which includes a glossary of terms to facilitate the sharing of information about biological diversity (Darwin Core Maintenance Group, 2021). Both of these aggregators streamline data access through application program interfaces (APIs)—specifically, through the GBIF Occurrence API and the iDigBio Search API, which can be accessed in R through their respective R packages, rgbif (Chamberlain et al., 2023) and ridigbio (Michonneau and Collins, 2022). iDigBio contains digitized specimen records primarily from U.S. collections, while GBIF contains both specimen and observation records found internationally. As of 23 February 2024, the iDigBio portal (https://www.idigbio.org/portal/search) contained more than 139 million specimen records, and the GBIF portal (https://www.gbif.org/occurrence/search) contained more than two billion occurrence records. Herbarium data downloaded from these and other aggregators have been used to investigate a diverse array of biological

questions (e.g., Wollan et al., 2008; Willis et al., 2017; Allen et al., 2019; reviewed in Heberling et al., 2019). Applications in research are broad and include species delimitation (e.g., Zapata and Jiménez, 2012; Su et al., 2020; de Mestier et al., 2023; Wu et al., 2023), understanding organismal response to seasonal events (e.g., Pearson et al., 2020; Guralnick et al., 2022; Belitz et al., 2023; Park et al., 2023), exploring global patterns of biodiversity (e.g., Gaynor et al., 2020; Melton et al., 2022; Folk et al., 2023), and investigating the potential impact of climate change on species distributions (e.g., Rawal et al., 2015; Gaynor et al., 2018; Hodel et al., 2022; Naranjo et al., 2022; Wang et al., 2022).

There are many R packages available for accessing biodiversity data from various online repositories, including galah (Westgate et al., 2023), bRacatus (Arlé et al., 2021), and plantR (de Lima et al., 2021). Although many of these R packages utilize rgbif or a custom wrapper to access the GBIF Occurrence API, none of these packages use ridigbio or the iDigBio Search API. There is currently only a single R package available to streamline data download from GBIF and iDigBio, spocc (Owens et al., 2023). However, spocc does not maximize the number of records returned from GBIF and iDigBio due to the search defaults (e.g., a low default download limit, exact matching, search methods) (see Methods and Results). Furthermore, complexities often arise when (1) trying to obtain all records corresponding to a single species due to the wide array of taxonomic identifiers (i.e., synonyms) that may exist, (2) identifying and removing problematic or arbitrary records, or (3) organizing the data in a readable fashion for downstream use.

To address these numerous challenges, we created gatoRs (Geographic And Taxonomic Occurrence R-based Scrubbing), an R package to help users navigate these critical data download and processing steps. We provide functions to streamline the processing of the data downloaded with our package; we also use interactive cleaning methods to provide users with greater control of the scrubbing process. These interactive methods have the added benefit of providing opportunities for educational demonstration in classroom settings.

Additionally, we provide a step-by-step workflow to help users employ this new package (Figure 1; Appendices S1, S2; see Supporting Information with this article). The gatoRs package is freely available at https://github.com/nataliepatten/gatoRs, as well as via The Comprehensive R Archive Network (CRAN), and can be installed, accessed, and used on any computer. gatoRs was created during the COVID-19 pandemic to facilitate research in botany and enable researchers and students to leverage readily available digitized biodiversity records by reducing the necessary programming requirements; while the functionality of this package facilitates research during times of limited mobility and/or resources, it is also of broad utility at any time.

Our aim was to create new tools for data acquisition in the form of software developments that were affordable and innovative. gatoRs includes functions that streamline downloading records from GBIF and iDigBio and takes into consideration traditional download differences between these aggregators, in contrast to spocc (Owens et al., 2023), the only other R package currently available to streamline downloads from both aggregators. We also developed a function that graphically displays flagged (i.e., potentially problematic) data points and allows these points to be manually reviewed and removed from the data set. Unlike other existing R packages, gatoRs provides

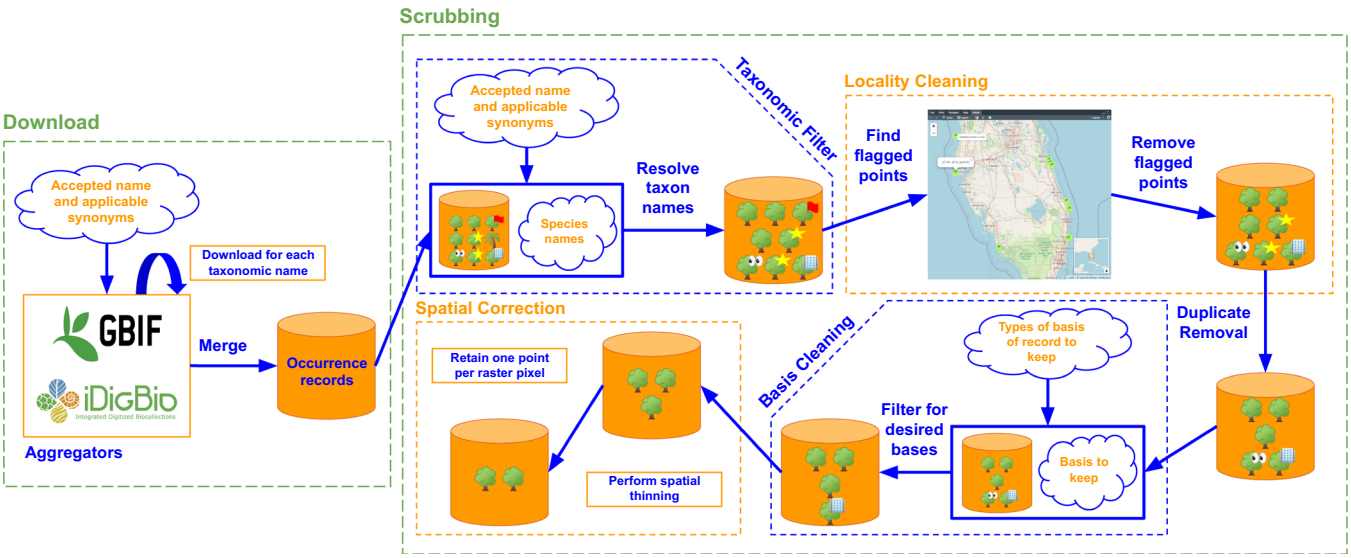

FIGURE 1 Schematic diagram of the gatoRs workflow. First, data are downloaded from GBIF and iDigBio for all applicable synonyms, and then the returned records are merged and returned. Second, scrubbing of data is completed through taxonomic harmonization, locality cleaning, duplicate removal, basis cleaning, and spatial correction.

an option for interactive cleaning, giving more control to the user over which points are deemed unfit. Additionally, we developed functions to identify records that need to be georeferenced and records with redacted information, as well as additional functions related to cleaning specimen records. The latter include functions to remove duplicate data points, resolve taxon names, and perform spatial correction. All of the functions (Table 1, Appendix S3) can be sorted into two categories: biodiversity data acquisition and occurrence record scrubbing.

METHODS AND RESULTS

Biodiversity data acquisition

In contrast to spocc, gatoRs has custom functions to accommodate the differences in download features and search logic between the GBIF Occurrence API and the iDigBio Search API (Appendix S3). spocc (Owens et al., 2023) uses nearly congruent record queries for both GBIF and iDigBio, despite these aggregators differing in the search logic associated with the DwC term scientificName. Specifically, iDigBio queries scientificName as an exact match, unlike GBIF, which allows a fuzzy match. Furthermore, Darwin Core indicates that the scientificName column should include authorities (https://dwc.tdwg.org/list/#dwc_scientificName); however, based on our observations, authorities are often omitted in the records published by data providers. This inconsistency is accounted for with GBIF's fuzzy match, which allows matches to records with and without authorities. Because the iDigBio API only returns exact matches when querying the scientificName field, a search of the scientific name with and without the authority is necessary to obtain all records associated with a specific species; otherwise only a subset of records for a given species is returned. Due to extensive spacing and spelling variability associated with scientific names, we found that two exact match queries are often not enough to identify all records associated with a species of interest. The solution we used to obtain all records for a species from iDigBio was to search all fields, and hence the entirety of each record, for a partial match of the search term(s). This approach returns a much greater number of records. However, these must be examined to avoid taxonomically inapplicable records (i.e., records that do not belong to a taxon of interest and therefore are erroneous records), which we filter within our function. To avoid taxonomically inapplicable records, we filter the records returned by our search of all fields to retain only those for which the scientificName column is found to be a fuzzy match (as defined by generalized Levenshtein edit distance for all gatoRs functions) to a value in the user-provided list containing the scientific name and all applicable synonyms. By using a fuzzy match, we avoid losing data that may or may not contain authorities, but remove inapplicable records, for example, where a species is mentioned only in the locality description.

Before records are available via the GBIF and iDigBio APIs, they are modified by both aggregators to improve data quality and remove data errors. However, automatic processing in any biodiversity data repository can lead to errors. For example, both GBIF and iDigBio allow records to be queried based on the date of collection; therefore, indexing of records in a database requires that associated collection year, month, and day information be present. When date information regarding month or day is missing, both GBIF and iDigBio will fill in these values with specified defaults; this leads to an overrepresentation of records for the first month of the year and for the first day of every month (Belitz et al., 2023). Given current ingestion and processing decisions and corresponding API search capabilities, we decided to utilize data providers' fields (i.e., the raw data) for iDigBio rather than the aggregator's fields. Furthermore, due to the computationally expensive methods required for downloading data provider fields for GBIF records, we provide an option to return these values but do not return them by default. Simple downloads for iDigBio and GBIF do not provide options to obtain the data supplied by the providers but instead return modified fields; hence, our workflow is distinct from the standard data download.

To resolve the problems described above, we created the function gators_download(), which streamlines the downloading process by downloading and returning records from iDigBio and GBIF independently based on a user-provided synonym list. A list of synonyms supplied to gators_download() should contain a species' accepted scientific name, as well as any scientific names considered a synonym for the species of interest. There are many tools available to construct synonym lists (reviewed in Grenié et al., 2023), for example, taxonstand (Cayuela et al., 2012), taxize (Chamberlain et al., 2020), WorldFlora (Kindt, 2020), and TNRS (Boyle et al., 2013; Maitner et al., 2023). However, taxonomic recommendations are extraneous to this package and are instead the user's purview, as gatoRs does not rely on any specific taxonomic backbone (see Appendix S2 for additional resources). Similar to spocc, records from GBIF and iDigBio are downloaded using functions from the rgbif (Chamberlain et al., 2023) and ridigbio (Michonneau and Collins, 2022) packages, respectively. However, unlike spocc, for iDigBio, data are downloaded by searching all fields for the provided species and selecting the provider columns, rather than the aggregator columns. In summary, gatoRs is unique in both how it searches iDigBio and which fields are obtained from iDigBio. Then, by default, to avoid taxonomically inapplicable records, we filter returned records to retain those with a fuzzy match to the scientific name based on a user-provided list as described above. For GBIF, data may be downloaded by searching either (1) the scientific name field with a fuzzy match or (2) the associated species key identified via the GBIF backbone taxonomy system. For records obtained from iDigBio and GBIF, we parse taxonomic information to ensure scientific name, genus, species, and infraspecific epithet are returned and replace

TABLE 1 An overview of the main functions provided in gatoRs, the required arguments for each function, and the general purposes of each function.

Category	Function[a]	Description
Data acquisition	gators_download(synonyms.list)	Downloads data from iDigBio and GBIF for a list of scientific names. Returns a data frame with 23 columns related to taxonomy, collection event, occurrence, storage, and collection location.
Identify missing data	need_to_georeference(df)	Identifies and returns records that are missing coordinate information but contain locality descriptions.
	needed_records(df)	Identifies and returns records that are redacted or withheld.
	remove_missing(df)	Removes records that are missing coordinate values and/or records that have been redacted or withheld.
	gators_merge(df1, df2)	Combines two data sets that have identical column names. Here, df1 and df2 should have the columns indicated for df.
Taxonomic harmonization	taxa_clean(df, synonyms.list)	Removes records based on the scientificName column. The type of filter used to remove records defaults to "fuzzy", but may be set to "exact" or "interactive". If an accepted.name argument is provided, the returned data frame will have an additional column titled accepted_name.
Locality cleaning	basic_locality_clean(df)	Removes records with missing coordinates, impossible coordinates, coordinates at (0,0), and where coordinates have been skewed. Precision of coordinates will be rounded to two decimal places by default.
	process_flagged(df)	Visualize and inspect occurrence records that may contain spatial errors with an interactive map. Based on console responses, records flagged as potential errors can be removed.
Remove duplicate records	remove_duplicates(df)	Remove specimen duplicates, aggregator duplicates, and within-aggregator duplicates based on each specimen's coordinates, occurrenceID, eventDate, ID, and aggregator.
Basis cleaning	basis_clean(df, basis.list)	Removes records based on the basisOfRecord column. If a basis.list is provided, records are filtered based on an exact match. If a basis.list is not provided, filtering is interactive.
Spatial correction	thin_points(df)	Removes records based on coordinate thinning with a minimum nearest neighbor distance approach. Minimum distance in kilometers can be specified with the argument distance.
	one_point_per_pixel(df)	Randomly samples and returns a data frame with a single occurrence record per raster pixel.
Cleaning wrapper	full_clean(df, synonyms.list)	Performs above cleaning steps with their default arguments. All cleaning steps, except taxonomic harmonization, can be bypassed by setting their associated arguments to FALSE.
Downstream data processing	data_chomp(df, accepted.name)	Returns data frame prepared for Maxent. This data frame has three columns: species, longitude, and latitude. Here, species is equal to the value indicated by accepted.name.
	citation_bellow(df)	Retrieves the citation information for records where aggregator = "GBIF". Returns a list with citation information for the GBIF data downloaded.
	remove_redacted(df)	Returns only records that were obtained from the aggregators GBIF and iDigBio, and thus only publicly available records. Removes any records that were obtained from other aggregators, which are presumably private or protected.

[a]Synonyms.list refers to a list object containing the accepted scientific name and all synonyms. df refers to a data frame created by the gators_download() function that contains 23 columns (scientificName, genus, specificEpithet, intraspecificEpithet, basisOfRecord, eventDate, year, month, day, occurrenceID, recordedBy, institutionCode, ID, informationWithheld, aggregator, country, county, stateProvince, locality, latitude, longitude, coordinateUncertaintyInMeters, and habitat). Unless otherwise indicated, the data frame supplied to a function will be returned with a modified number of rows, but the number and identity of the columns will remain the same.

empty observations when information is known. Our function also queries the species names in the data set and fixes incorrect capitalization, following standard capitalization of scientific names. Common capitalization errors include lack of any capitalization, incorrect capitalization of subspecies and variety, and lack of capitalization of the authority name(s). After merging records from both data aggregators, we retain columns related to taxonomy (scientificName, genus, specificEpithet, intraspecificEpithet), collection event (basisOfRecord, eventDate, year, month, day), occurrence (occurrenceID, recordedBy), storage (institutionCode, ID, informationWithheld, aggregator), and location (country, county, stateProvince,

locality, latitude, longitude, coordinateUncertaintyInMeters, habitat). When a filename is provided, this function will automatically create a .csv file for the downloaded data. We highly recommend that users save a copy of their downloaded data prior to any modification, as all subsequent processing steps within our package will produce data sets with records removed rather than flagged.

Occurrence record scrubbing

Before using digitized biodiversity records to explore biological questions, the records must be further vetted and filtered to remove records that are not appropriate for use in downstream research aims because of errors, incompleteness, or both. We designed additional functions to streamline the identification of missing data, taxonomic harmonization, locality filtering, and duplicate correction, decrease spatial bias, and filter records according to the basisOfRecord (e.g., PreservedSpecimen, HumanObservation) (Table 1, Appendix S3). Finally, we streamlined the cleaning and scrubbing processes into a single function. Unless otherwise stated, all scrubbing functions remove records, or rows, but retain all columns created by the gator_download() function (scientificName, genus, specificEpithet, intraspecificEpithet, basisOfRecord, eventDate, year, month, day, occurrenceID, recordedBy, institutionCode, ID, informationWithheld, aggregator, country, county, stateProvince, locality, latitude, longitude, coordinateUncertaintyInMeters, habitat). Additionally, certain scrubbing functions allow users to respond in the console while processing a data set; we refer to this feature as interactive. Although interactive approaches to data scrubbing may be time consuming, filtering a data set based on an exact match may remove applicable data, while a fuzzy match may retain erroneous records. Notably, additional R packages focused on scrubbing biodiversity data for research purposes are available, for example, bdc (Ribeiro et al., 2023) and CoordinateCleaner (Zizka et al., 2019). Many graphical user interface (GUI)–based R packages for cleaning biodiversity data exist, including, but not limited to, bdchecks (Gibas et al., 2019), bdclean (Nagarajah et al., 2019), and wallace (Kass et al., 2023).

Identifying missing locality data

Before processing and filtering records, users may collect additional data by obtaining or requesting coordinate values for geographically sparse taxa. Occasionally, records will have locality information represented only as qualitative descriptions, rather than quantitative latitude/longitude GPS coordinates. These records will need to be manually georeferenced if coordinates are required for downstream processing. Georeferencing is the process of taking a locality description and converting it to numerical coordinates (Wieczorek et al., 2004; Hackeloeer et al., 2014; Yao, 2020). To subset records needing georeferencing, we created a function called need_to_georeference(), which identifies records that lack coordinates and contain locality strings. Additionally, when the species of interest is listed as endangered or threatened, locality information may be redacted by the data provider. To obtain redacted locality information, users will need to contact the herbarium or collection directly and request the missing information for use in research. To identify records for which locality information has been redacted, we designed the function needed_records(), which subsets the records containing flags indicating the locality information has been redacted. Based on the subsetted data frame of redacted records, the herbarium to contact can be identified based on the associated collection code. To identify contact information based on a collection code, users should refer to Index Herbariorum (http://sweetgum.nybg.org/science/ih/).

After obtaining additional locality information for specimen records, a user should merge the missing data with the main data frame prior to any additional processing steps. To merge the two data frames, the user must format the acquired records in the same format as the data frame obtained with gators_download() and should indicate the source of the records in the aggregator column. We provide the function gators_merge() to aid in merging two data sets with identical column names. However, often column names and values may not match; therefore, we suggest merging of the initial data sets and any obtained records using the bdc package's function, bdc_standardize_datasets() (Ribeiro et al., 2023). Prior to merging any obtained data with the original data frame, users should remove records identified with the need_to_georeference() and needed_records() functions using the remove_missing() function to avoid duplicating records. For any redacted records obtained, as mentioned above, users should indicate the source of the record in the aggregator column. By indicating an aggregator not equal to iDigBio or GBIF, these records can easily be removed with the remove_redacted() function to prevent any accidental publication of these records.

Taxonomic harmonization

After all data are obtained, the first step in data cleaning is taxonomic harmonization. Taxonomic errors can negatively impact research and its applications (e.g., Daugherty et al., 1990; Jin and Yang, 2020). Specimen records often lack proper and accepted scientific name designations due to the time-consuming nature of updating scientific names for specimen records. Users must address the lack of taxon harmonization in downloaded records for downstream applications; therefore, it is crucial that users are knowledgeable about each species of interest, its synonyms, and the currently accepted scientific name. To harmonize taxonomic identifiers for downloaded specimen records, we designed a function titled taxa_clean(), which filters

records based on either an exact or fuzzy match to a supplied synonym list, as well as allows users to employ an interactive approach. Due to variation in spelling and potential inclusion of author strings, an exact match may remove applicable records, and a fuzzy match may not remove all erroneous records. Therefore, to obtain the most taxonomically applicable records and remove taxonomically erroneous records, we suggest an interactive approach. If users select the interactive approach, the function will first print all unique scientific names in the current data set and then ask the user to respond in the console to prompts regarding which records, if any, should be removed based on their scientific name. After filtering, based on a user-provided taxonomy, an accepted name column can be defined using an optional argument.

Locality cleaning

Locality information is often vital for downstream uses in biodiversity research (e.g., Ritter et al., 2019; Kass et al., 2022). Here, we provide the basic_locality_clean() function to remove records with missing coordinates (i.e., lacking latitude and/or longitude) and remove impossible coordinates (greater than 180 or less than −180 for longitude; greater than 90 or less than −90 for latitude). The basic_locality_clean() function can be used to remove records found at the intersection of the equator and prime meridian (latitude and longitude of 0) and to remove coordinates skewed due to a species' protected status. Lastly, this function can be used to round values to a desired precision.

In addition to a basic locality filter, records should be filtered to remove spatial outliers. Spatial outliers may exist due to human errors in data recording or in species identification. Additionally, records may be from locations outside of a species' range because they represent cultivated individuals, often grown in botanical gardens. We use the CoordinateCleaner package (Zizka et al., 2019) to identify occurrence records that may contain spatial errors or flagged records. Specifically, CoordinateCleaner's clean_coordinates() function identifies and flags records that are located in state capitals, country centroids, the GBIF headquarters, biodiversity institutions (including botanical gardens, museums, herbaria, etc.), and the ocean (Zizka et al., 2019). It also flags records with equal latitude and longitude, records with precisely zero coordinates, and records with locality outliers (Zizka et al., 2019). Because cultivated or erroneous records may still be missed by this approach, visual inspection of occurrence records is recommended. We streamline identification of outliers with an interactive function, process_flagged(), where users can visualize both flagged and non-flagged records, labeled in red and blue, respectively, to determine which points, if any, should be removed. The map visualization also provides details regarding the reason each point was flagged, such as geographic outlier, sea coordinates, or equal coordinates. This function gives control to the user in determining the validity of the data; the user can choose to remove all the problematic points, or none, depending on the research purposes and downstream applications.

Removing duplicate records

When using herbarium specimens to document occurrences, two types of duplicate records may be present: (1) specimen duplicates and (2) aggregator duplicates. A specimen duplicate is when two or more specimens represent a single gathering of a single species (or infraspecific taxon), often from the same individual plant, made by a collector at one time. Depositing specimen duplicates to multiple herbaria is extremely common; for example, a recent study identified that when downloading records from GBIF, on average over 30% of the records for a taxon were duplicates (Zizka et al., 2020). These specimen duplicates share a specimen-level identifier (dwc:occurrenceID; http://rs.tdwg.org/dwc/terms/occurrenceID) that is persistent, globally unique, and identifies an occurrence of an organism (dwc:Organism; http://rs.tdwg.org/dwc/terms/Organism) at a specific location and time (Nelson et al., 2018; Mabry et al., 2022). The occurrenceID can therefore be used to identify and remove specimen duplicates from a data set to prevent inflation in the number of occurrences. However, occurrenceID is not always provided to data aggregators, making recognition of specimen duplicates difficult. Aggregator duplicates exist when a single record is indexed by multiple aggregators and are artifacts that must be removed. Both specimen and aggregator duplicates should be identifiable by the occurrenceID (if available), and both can also be recognized based on coordinate values and date of collection. For example, if two or more records share the same collection date and the same location, these records may be considered duplicates for the purpose of assembling a list of occurrences, and duplicates can be removed. We note that records for the same taxon with identical coordinates and date of collection may actually represent distinct individual plants (with distinct occurrenceIDs that were not reported to the aggregators) and may be an important source of data for some applications; in this case, a user may choose not to consider these records as duplicates but to retain them for further consideration.

Our remove_duplicates() function removes both specimen and aggregator duplicates, as well as within-aggregator duplicates that may accumulate due to processing errors. Aggregators assign unique identifiers—GBIF assigns keys, and iDigBio assigns universally unique identifiers (UUID)—and we leverage these identifiers to remove any duplicates that may exist within each aggregator. We then use coordinates, occurrenceID, and eventDate to identify and remove specimen and aggregator duplicates. To leverage all date information available, we populate the year, month, and day columns (if not already provided) using the eventDate column. To parse eventDate, we attempt the ISO 8601 parsing methods from the parsedate package with their functions

parse_iso_8601() and format_iso_8601() (Csárdi and Torvalds, 2022). Notably, ISO 8601 only includes time since the Unix epoch, or 1 January 1970; therefore, dates that occur before 1970 may not be automatically parsed. If we are unable to parse the included date for particular records, two options are provided: (1) automatically remove the few unparsable dates or (2) manually enter the year, month, and day for these records when prompted. If the user chooses to manually enter the event date, the record's eventDate will be printed, and the user will be asked to manually enter the year, month, and day of this eventDate into the console. Users are only prompted to manually parse event dates for records where year, month, and day are absent, but eventDate is present and cannot be parsed. Based on the prompt printed with this function, a user may not be able to identify the record for which the event is being parsed. Finally, we filter rows to only return records with distinct latitude, longitude, occurrenceID, year, month, and day.

Basis cleaning

In some instances, users may want to retain only records associated with physical specimens due to common misidentification (McDonough MacKenzie et al., 2017). Alternatively, a user may be interested in exploring trends regarding community observations (Grade et al., 2022; White et al., 2023). The basis_clean() function offers both an interactive and automated method, using a fuzzy match, to filter for basisOfRecord, as the user may wish to remove certain types of records. Examples of basisOfRecord include PreservedSpecimen, FossilSpecimen, and HumanObservation (Darwin Core Maintenance Group, 2021). As mentioned above, fuzzy matches may retain erroneous records; therefore, an interactive approach may be preferred. With the interactive method, the function will print all unique basisOfRecord values in the current data set and then ask the user to respond in the console to prompts regarding which records, if any, should be removed based on their basisOfRecord. Alternatively, the user can input a list of types of records to retain from the data set, and records with a different basisOfRecord will be removed automatically.

Spatial correction

Collection efforts often lead to geographic clustering of specimens due to accessibility and infrastructure, as observed for both herbarium (Daru et al., 2018) and community science (Steen et al., 2021; Grade et al., 2022) efforts. The majority of specimen collections occur in high-traffic, urban areas due to the physical barriers associated with reaching more remote areas, resulting in geographic bias (Meineke and Daru, 2021). In particular, recent large-scale production of unvouchered specimens has not increased geographic data coverage; rather, there is a bias toward regions with easily accessible sampling sites (Daru and Rodriguez, 2023). In addition, geographic bias can occur due to low sampling intensity in some areas with high biodiversity (Meineke and Daru, 2021). Spatial clustering can lead to incorrect interpretations of a species' current range, inflate confidence in ecological niche estimates (Veloz, 2009), and influence the quality of species distribution models (Kramer-Schadt et al., 2013; Aiello-Lammens et al., 2015; Kiedrzyński et al., 2017; Steen et al., 2021). To reduce geographic bias prior to downstream applications, users often employ spatial thinning.

Hence, we provide the thin_points() function to perform spatial thinning using the spThin package (Aiello-Lammens et al., 2015) by reducing records based on a minimum nearest neighbor distance (NND) approach. The thinning algorithm provided by spThin calculates the pairwise distances between data points, identifies the number of neighboring points within the minimum NND, and then samples and removes a single record randomly (Aiello-Lammens et al., 2015). Points within the NND distances are repeatedly identified and removed until all records have met the minimum NND requirement (Aiello-Lammens et al., 2015).

When modeling the fundamental niche of a species, presence-absence models only consider one point per geographic pixel (Phillips, 2017); therefore, users often reduce their data set to one point per pixel prior to using the records to explore climatic niche. For this purpose, we created a simple function, one_point_per_pixel(), to reduce the number of records to only one point per raster pixel based on the maximum resolution of an inputted geographic raster object.

Occurrence data cleaning overview

Herbaria serve as large data sources for plant functional traits in exploration of important biological questions in research areas such as water-use efficiency, plant–pollinator interactions, plant hyperaccumulation, and functional group adaptations (Heberling, 2022). Large-scale exploration of biodiversity data enables new hypotheses and discoveries in evolutionary biology, such as the potential relationship between phylogenetic diversity and phenotypic evolution in an area (Soltis and Soltis, 2016). Biodiversity "big data" also are used to classify habitats, such as forests (Agrillo et al., 2021). "Big data" management strategies can be used to improve the quality of biological data provided by community science platforms such as eBird for research applications (Kelling et al., 2015).

To streamline data processing, our full_clean() function automates vetting and filtering by wrapping all cleaning functions into a single step. This function was designed to optimize data processing so beginning programmers can process their occurrence record data and experienced programmers can expedite the time-consuming cleaning process required before proceeding to downstream research applications. This function is entirely automated and thus does not take advantage of the interactive options provided in the individual cleaning functions. Using this wrapper is recommended for data processing that does not require

interactive/manual cleaning and inspection, or on large data sets where this would be time consuming. All cleaning steps, except taxonomic harmonization, can be bypassed by setting their associated argument to FALSE.

Downstream data processing

We created three functions to aid users in data export and preparation of publications. Regarding data export, many users may want to employ Maxent (Phillips et al., 2006) for species distribution modeling; therefore, our function data_chomp() subsets the data set to include only the columns needed for this downstream application, i.e., the user-provided accepted name and coordinate value columns. To aid in data preparation for publication and to comply with GBIF's data use agreement, our citation_bellow() function will return the citation information for these records as a list (this function name is based on alligators bellowing). Finally, remove_redacted() will remove records where the aggregator value is not equal to iDigBio or GBIF (see "Identifying missing locality data").

Comparison of data downloads using gatoRs and spocc

We compared the number of records obtained from our gators_download() function to the occ() function from spocc (Owens et al., 2023) for 25 plant species downloaded on 4 August 2023. We defined the synonym list for each species based on a careful literature review. Due to the download limits set by both gatoRs and spocc, the default gators_download() function would seem to require more computational time than the occ() function; however, when download limits are equal, the two functions have similar computation times. The default download function from spocc only allows 500 records from each source, while gatoRs allows 100,000 records from each source. Even if spocc's limit is modified to equal 100,000, we found that the gators_download() function obtained more records for all species except *Polygonum basiramia* (Small) T. M. Schust. & Reveal (Figure 2, Appendix S4). For all 25 species, gatoRs obtained more records from iDigBio than spocc for both download limits (Figure 2, Appendix S4). For some species, gatoRs retrieved fewer records from GBIF than retrieved with spocc. However,

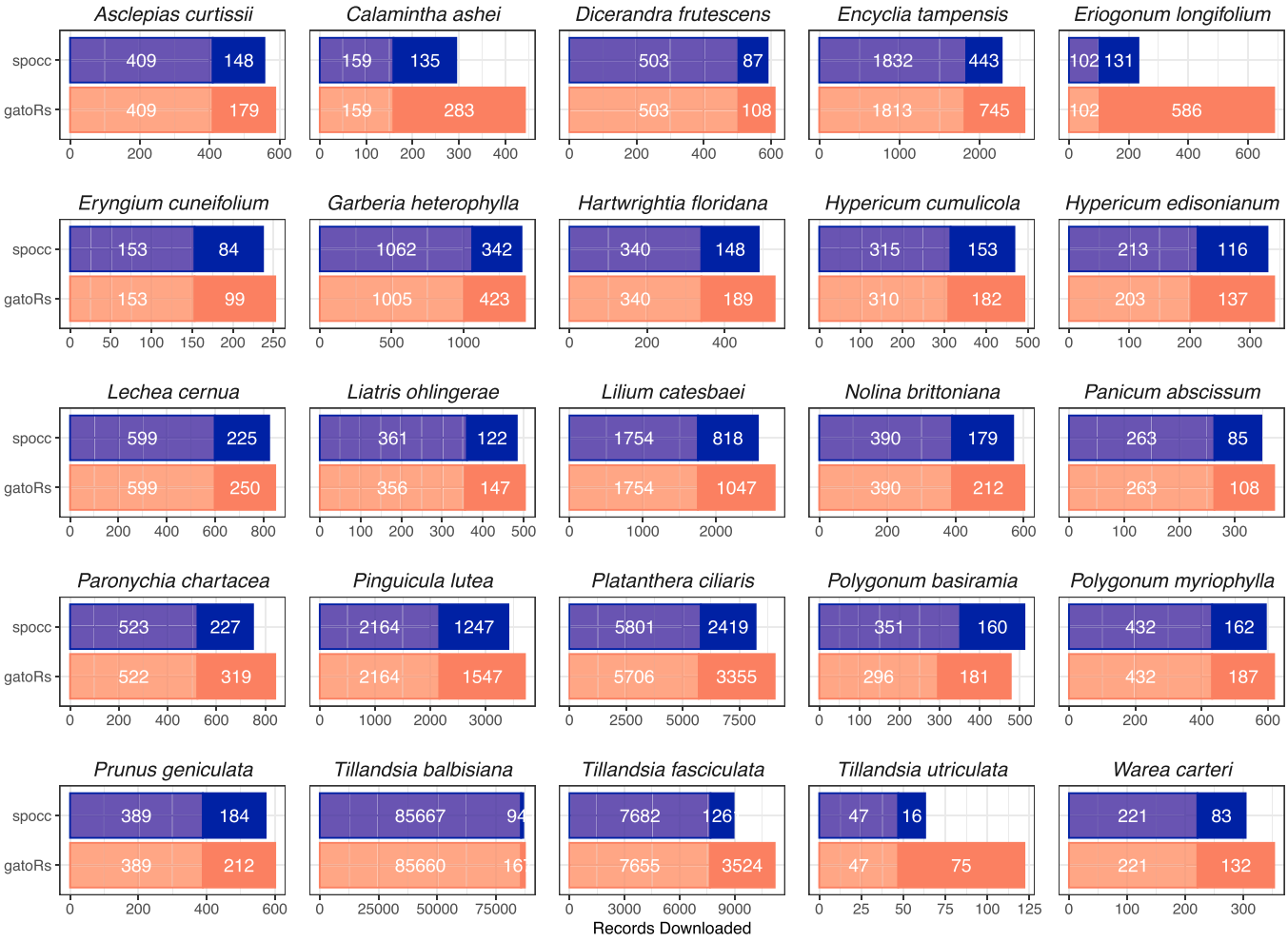

FIGURE 2 Number of records downloaded from spocc (blue; limit = 100,000) and gatoRs (orange) from GBIF (light shade) and iDigBio (dark shade). Note, *Eriogonum longifolium* var. *gnaphalifolium* is shortened to *Eriogonum longifolium* in the figure.

in all cases, the higher number of GBIF records retained by spocc was simply due to inclusion of duplicate records, i.e., records with the same key or UUID (Appendix S5). Once we removed only within-aggregator duplicates with the distinct function from dplyr (Wickham et al., 2023), we found that the number of records retained from GBIF by gatoRs and spocc (limit = 100,000) was equal for all 25 species (Appendix S5). Additionally, gatoRs retrieved more records for *Polygonum basiramia* than spocc after within-aggregator duplicates were removed (Appendix S5).

To assess the quality of the data retrieved by gatoRs compared to spocc, we applied the taxa_clean(), basic_locality_clean(), and remove_duplicates() functions to both downloads (Figure 3, Appendix S6) and found that the gatoRs data set included more records for all species. After data scrubbing, only 12% of the total records remained for the spocc data (limit = 100,000) for all 25 species, while about 78% of the records remained for the data set downloaded with gatoRs (Figure 3). The largest percentage of records removed during the data scrubbing processes was removed with basic_locality_clean() for gatoRs, as well as for spocc when the download limit is set to 500. When the download limit is set to 100,000 for spocc, the largest percentage of removed records occurred with the remove_duplicates() step, which is likely due to the lack of occurrenceID for these records and the retention of within-aggregator duplicates during the download step (Appendices S4, S5). Overall, after each scrubbing step, the gatoRs data set had more records than the spocc data set for both spocc download limits. The difference between the number of records retained after data scrubbing from the gatoRs and spocc data sets was modest for some taxa and extreme for others; large differences were noted for *Eriogonum longifolium* Nutt. var. *gnaphalifolium* Gand. (182 additional records), *Pinguicula lutea* Walter (57 additional records), *Platanthera ciliaris* (L.) Lindl. (125 additional records), *Tillandsia balbisiana* Schult. f. (81,772 additional records), and *Tillandsia fasciculata* Sw. (1882 additional records) (Appendix S5). Hence, the gatoRs package provides many more occurrence records for use in various applications, an important advantage when the number of records increases the accuracy of the results, such as in species distribution modeling (van Proosdij et al., 2016).

Additionally, for some species, GBIF provided the majority of records (this is especially clear for *Tillandsia*

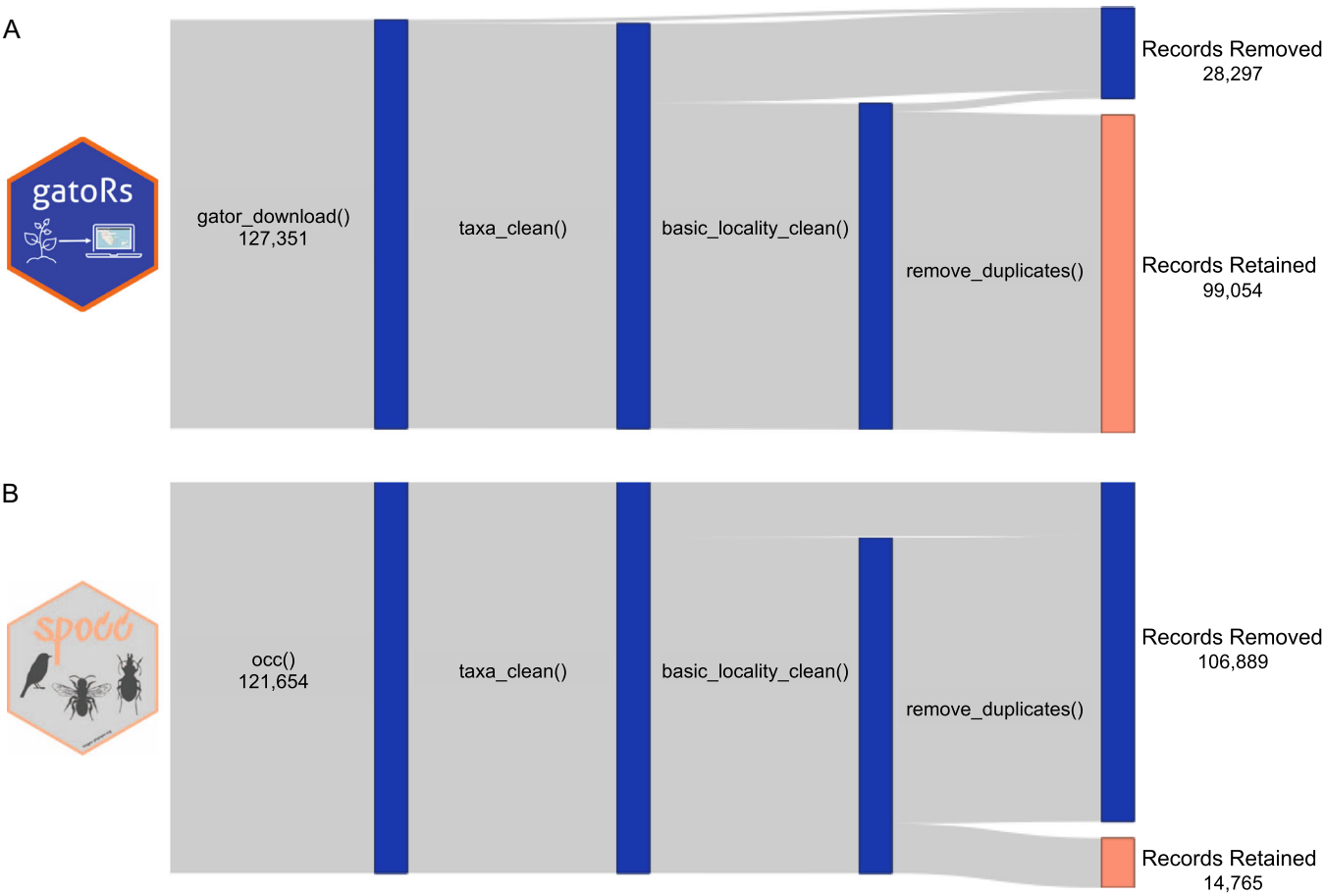

FIGURE 3 Sankey diagrams showing the sum of records returned for all 25 species after each cleaning step when using (A) gators_download() from gatoRs and (B) occ() from spocc with the limit set to 100,000. This Sankey diagram was generated using the networkD3 R package (Allaire et al., 2017) and was inspired by Panter et al. (2020) (see their Figure 3). The number of records after each processing step can be found in Appendix S6. The spocc logo was sourced from https://github.com/ropensci/spocc/blob/master/man/figures/logo.png.

balbisiana), whereas for others, more records were available from iDigBio (in particular, for *Eriogonum longifolium* var. *gnaphalifolium*) (Figure 2). Thus, retrieving fields from both aggregators allows for a greater number of records to be obtained—even after removing the duplicate records recorded by both GBIF and iDigBio.

Based on these cleaned records, we defined minimum convex hulls, trimmed to only include land mass, for each species. We found the additional records obtained by gatoRs led to increased range estimations for *Calamintha ashei* (Weath.) Shinners, *Eriogonum longifolium* var. *gnaphalifolium*, *Paronychia chartacea* Fernald, *Tillandsia balbisiana*, and *Tillandsia utriculata* L. (Appendix S7).

CONCLUSIONS

We have shown that gatoRs streamlines the downloading and scrubbing of biodiversity data from GBIF and iDigBio. As a result, it is especially useful for users who have limited familiarity with R or programming in general. Furthermore, our package addresses differences in search logic between GBIF and iDigBio, a feature that is unavailable with spocc or any other packages for downloading occurrences from both aggregators, and uses methods to download the maximum number of records for a taxon of interest. We anticipate that these features of gatoRs will be useful for research applications that require a minimum number of records to provide useful and accurate predictions. This utility is especially true for endangered, threatened, or otherwise rare species where data are limited. In these cases, it is crucial to download all records associated with these species to enable accurate research applications (e.g., Panter et al., 2020). As noted, by obtaining additional records with gatoRs, range size estimates increased for some of our species of interest (Appendix S7). Our package also incorporates unique interactive capability for cleaning data, unlike other available R packages. Because of this software functionality, gatoRs enables data cleaning with user control, which differs from other currently available methods. In addition, we provide a simple cleaning wrapper function that performs essential cleaning processes all in one step, with no user input required. This feature streamlines the cleaning process and is a valuable asset for iteratively cleaning multiple data sets, allowing users to focus on the plethora of downstream applications related to specimen data. Additional sources of error and bias are not addressed in gatoRs (e.g., collector bias; Baldwin et al., 2017), and such issues could be implemented in the future. Overall, gatoRs provides greater access to occurrence records and thus facilitates discussion of herbarium specimens and their potential for biodiversity-related research. In this way, our package has the potential to provide important educational benefits in the form of hands-on teaching with demonstrations of downloading and cleaning data through the various gatoRs functions, especially when taking advantage of the interactive options.

We believe that our free and widely available R package will greatly increase access and usage of herbarium data for both advanced researchers and beginning students. Finally, although we illustrate the use of gatoRs with herbarium specimen data, it is not restricted to herbarium data but can be used for any basis of record available via GBIF and iDigBio; it will promote further research using biodiversity data and stimulate increased student interest in the study of biodiversity.

AUTHOR CONTRIBUTIONS

N.N.P. and M.L.G. programmed the R package, with suggestions from P.S.S. and D.E.S. N.N.P. and M.L.G. wrote the initial manuscript draft, and all authors contributed to revising and editing the final manuscript. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

This project emerged from a CURE (Course-based Undergraduate Research Experience) course taught by P.S.S., D.E.S., and M.L.G. at the University of Florida (BOT 2930) with support from the UF Center for Undergraduate Research. This research was supported by National Science Foundation (NSF) grants DBI-1547229 and DBI-2027654 to P.S.S. and by an NSF Graduate Research Fellowship (DGE-1842473) to M.L.G. N.N.P. was supported by the University of Florida University Research Scholars Program, University Scholars Program, and Emerging Scholars Program. We thank J. T. Miller, L. E. Gillett, and M. E. Mabry (University of Florida) for reviewing and testing this package, and M. W. Belitz (Michigan State University) for reviewing the code. This package will be actively maintained under DBI-2027654.

DATA AVAILABILITY STATEMENT

The gatoRs R package source code is freely and publicly available at https://github.com/nataliepatten/gatoRs and on CRAN (https://CRAN.R-project.org/package=gatoRs), and our user guide can be accessed at https://nataliepatten.github.io/gatoRs/. Code and data associated with the download comparison and generation of Figures 2 and 3 are available at https://zenodo.org/record/8326603. Additional code examples can be found at https://github.com/soltislab/BotanyENMWorkshops.

ORCID

Natalie N. Patten http://orcid.org/0000-0001-8090-1324
Michelle L. Gaynor http://orcid.org/0000-0002-3912-6079
Douglas E. Soltis http://orcid.org/0000-0001-8638-4137
Pamela S. Soltis http://orcid.org/0000-0001-9310-8659

REFERENCES

Agrillo, E., F. Filipponi, A. Pezzarossa, L. Casella, D. Smiraglia, A. Orasi, F. Attorre, and A. Taramelli. 2021. Earth observation and biodiversity big data for forest habitat types classification and mapping. *Remote Sensing* 13(7): 1231.

Aiello-Lammens, M. E., R. A. Boria, A. Radosavljevic, B. Vilela, and R. P. Anderson. 2015. spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography* 38: 541–545.

Allaire, J. J., C. Gandrud, K. Russell, and C. J. Yetman. 2017. networkD3: D3 JavaScript network graphs from R. Website: https://CRAN.R-project.org/package=networkD3 [accessed 13 February 2024].

Allen, J. M., R. A. Folk, P. S. Soltis, D. E. Soltis, and R. P. Guralnick. 2019. Biodiversity synthesis across the green branches of the tree of life. *Nature Plants* 5(1): 11–13.

Arlé, E., A. Zizka, P. Keil, M. Winter, F. Essl, T. Knight, P. Weigelt, et al. 2021. bRacatus: A method to estimate the accuracy and biogeographical status of georeferenced biological data. *Methods in Ecology and Evolution* 12(9): 1609–1619.

Bakker, F. T., A. Antonelli, J. A. Clarke, J. A. Cook, S. V. Edwards, P. G. P. Ericson, S. Faurby, et al. 2020. The global museum: Natural history collections and the future of evolutionary science and public education. *PeerJ* 8: e8225.

Baldwin, B. G., A. H. Thornhill, W. A. Freyman, D. D. Ackerly, M. M. Kling, N. Morueta-Holme, and B. D. Mishler. 2017. Species richness and endemism in the native flora of California. *American Journal of Botany* 104(3): 487–501.

Belitz, M. W., E. A. Larsen, V. Shirey, D. Li, and R. P. Guralnick. 2023. Phenological research based on natural history collections: Practical guidelines and a lepidopteran case study. *Functional Ecology* 37(2): 234–247.

Boyle, B., N. Hopkins, Z. Lu, J. A. Raygoza Garay, D. Mozzherin, T. Rees, N. Matasci, et al. 2013. The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinformatics* 14(16): 16.

Cayuela, L., Í. Granzow-de la Cerda, F. S. Albuquerque, and D. J. Golicher. 2012. taxonstand: An R package for species names standardisation in vegetation databases. *Methods in Ecology and Evolution* 3(6): 1078–1083.

Chamberlain, S., E. Szoecs, Z. Foster, Z. Arendsee, C. Boettiger, K. Ram, I. Bartomeus, et al. 2020. taxize: Taxonomic information from around the web. Website: https://github.com/ropensci/taxize [accessed 13 February 2024].

Chamberlain, S., V. Barve, D. Mcglinn, D. Oldoni, P. Desmet, L. Geffert, and K. Ram. 2023. rgbif: Interface to the Global Biodiversity Information Facility API. Website: https://CRAN.R-project.org/package=rgbif [accessed 13 February 2024].

Csárdi, G., and L. Torvalds. 2022. parsedate: Recognize and parse dates in various formats, including all ISO 8601 formats. Website: https://CRAN.R-project.org/package=parsedate [accessed 13 February 2024].

Daru, B. H., D. S. Park, R. B. Primack, C. G. Willis, D. S. Barrington, T. J. S. Whitfeld, T. G. Seidler, et al. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist* 217(2): 939–955.

Daru, B. H., and J. Rodriguez. 2023. Mass production of unvouchered records fails to represent global biodiversity patterns. *Nature Ecology & Evolution* 7: 816–831.

Darwin Core Maintenance Group. 2021. List of Darwin Core terms. Website: https://dwc.tdwg.org/list/ [accessed 13 February 2024].

Daugherty, C. H., A. Cree, J. M. Hay, and M. B. Thompson. 1990. Neglected taxonomy and continuing extinctions of tuatara (*Sphenodon*). *Nature* 347: 177–179.

de Lima, R. A. F., A. Sánchez-Tapia, S. R. Mortara, H. ter Steege, and M. F. de Siqueira. 2021. plantR: An R package and workflow for managing species records from biological collections. *Methods in Ecology and Evolution* 14(2): 332–339.

de Mestier, A., R. Lücking, J. Gutierrez, G. Brokamp, M. Celis, and T. Borsch. 2023. Nested singletons in molecular trees: Utility of adding morphological and geographical data from digitized herbarium specimens to test taxon concepts at species level in the case of *Casearia* (Salicaceae). *Ecology and Evolution* 13(1): e9736.

Folk, R. A., M. L. Gaynor, N. J. Engle-Wrye, B. C. O'Meara, P. S. Soltis, D. E. Soltis, R. P. Guralnick, et al. 2023. Identifying climatic drivers of hybridization with a new ancestral niche reconstruction method. *Systematic Biology* 72: 856–873.

Gaynor, M. L., D. B. Marchant, D. E. Soltis, and P. S. Soltis. 2018. Climatic niche comparison among ploidal levels in the classic autopolyploid system, *Galax urceolata*. *American Journal of Botany* 105(10): 1631–1642.

Gaynor, M. L., C.-N. Fu, L.-M. Gao, L.-M. Lu, D. E. Soltis, and P. S. Soltis. 2020. Biogeography and ecological niche evolution in Diapensiaceae inferred from phylogenetic analysis. *Journal of Systematics and Evolution* 58(5): 646–662.

Gibas, P., T. Gueta, V. Barve, T. Nagarajah, and Y. Carmel. 2019. bdchecks: Biodiversity Data Checks. Website: https://CRAN.R-project.org/package=bdchecks [accessed 13 February 2024].

Grade, A. M., N. W. Chan, P. Gajbhiye, D. J. Perkins, and P. S. Warren. 2022. Evaluating the use of semi-structured crowdsourced data to quantify inequitable access to urban biodiversity: A case study with eBird. *PLoS ONE* 17(11): e0277223.

Grenié, M., E. Berti, J. Carvajal-Quintero, G. M. L. Dädlow, A. Sagouis, and M. Winter. 2023. Harmonizing taxon names in biodiversity data: A review of tools, databases, and best practices. *Methods in Ecology and Evolution* 14(1): 12–25.

Guralnick, R., L. Campbell, and M. Belitz. 2022. Weather anomalies more important than climate means in driving insect phenology. *Communications Biology* 6: 490.

Hackeloeer, A., K. Klasing, J. M. Krisp, and L. Meng. 2014. Georeferencing: A review of methods and applications. *Annals of GIS* 20(1): 61–69.

Heberling, J. M. 2022. Herbaria as big data sources of plant traits. *International Journal of Plant Sciences* 183(2): 87–118.

Heberling, J. M., L. A. Prather, and S. J. Tonsor. 2019. The changing uses of herbarium data in an era of global change: An overview using automated content analysis. *BioScience* 69(10): 812–822.

Hodel, R. G. J., D. E. Soltis, and P. S. Soltis. 2022. Hindcast-validated species distribution models reveal future vulnerabilities of mangroves and salt marsh species. *Ecology and Evolution* 12(9): e9252.

Jin, J., and J. Yang. 2020. BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. *Global Ecology and Conservation* 21: e00852.

Kass, J. M., B. Guénard, K. L. Dudley, C. N. Jenkins, F. Azuma, B. L. Fisher, C. L. Parr, et al. 2022. The global distribution of known and undiscovered ant biodiversity. *Science Advances* 8(31): eabp9908.

Kass, J. M., G. E. Pinilla-Buitrago, A. Paz, B. A. Johnson, V. Grisales-Betancur, S. I. Meenan, D. Attali, et al. 2023. wallace 2: A shiny app for modeling species niches and distributions redesigned to facilitate expansion via module contributions. *Ecography* 2023(3): e06547.

Kelling, S., D. Fink, F. A. La Sorte, A. Johnston, N. E. Bruns, and W. M. Hochachka. 2015. Taking a 'Big Data' approach to data quality in a citizen science project. *Ambio* 44: 601–611.

Kiedrzyński, M., K. M. Zielińska, A. Rewicz, and E. Kiedrzyńska. 2017. Habitat and spatial thinning improve the Maxent models performed with incomplete data. *Journal of Geophysical Research: Biogeosciences* 122(6): 1359–1370.

Kindt, R. 2020. WorldFlora: An R package for exact and fuzzy matching of plant names against the World Flora Online taxonomic backbone data. *Applications in Plant Sciences* 8(9): e11388.

Kramer-Schadt, S., J. Niedballa, J. D. Pilgrim, B. Schröder, J. Lindenborn, V. Reinfelder, M. Stillfried, et al. 2013. The importance of correcting for sampling bias in Maxent species distribution models. *Diversity and Distributions* 19(11): 1366–1379.

Mabry, M. E., F. Zapata, D. L. Paul, P. M. O'Connor, P. S. Soltis, D. C. Blackburn, and N. B. Simmons. 2022. Monographs as a nexus for building extended specimen networks using persistent identifiers. *Bulletin of the Society of Systematic Biologists* 1(1): 8323.

Maitner, B., B. Boyle, and P. Efren. 2023. TNRS: Taxonomic Name Resolution Service. Website: https://CRAN.R-project.org/package=TNRS [accessed 13 February 2024].

McDonough MacKenzie, C., G. Murray, R. Primack, and D. Weihrauch. 2017. Lessons from citizen science: Assessing volunteer-collected plant phenology data with Mountain Watch. *Biological Conservation* 208: 121–126.

Meineke, E. K., and B. H. Daru. 2021. Bias assessments to expand research harnessing biological collections. *Trends in Ecology & Evolution* 36(12): 1071–1082.

Melton, A. E., M. H. Clinton, D. N. Wasoff, L. Lu, H. Hu, Z. Chen, K. Ma, et al. 2022. Climatic niche comparisons of eastern North American and eastern Asian disjunct plant genera. *Global Ecology and Biogeography* 31(7): 1290–1302.

Michonneau, F., and M. Collins. 2022. ridigbio: Interface to the iDigBio Data API. Website: https://CRAN.R-project.org/package=ridigbio [accessed 13 February 2024].

Nagarajah, T., T. Gueta, V. Barve, A. Agrawal, P. Gibas, and Y. Carmel. 2019. bdclean: A user-friendly biodiversity data cleaning app for the inexperienced R user. Website: https://CRAN.R-project.org/package=bdclean [accessed 13 February 2024].

Naranjo, A. A., A. E. Melton, D. E. Soltis, and P. S. Soltis. 2022. Endemism, projected climate change, and identifying species of critical concern in the Scrub Mint clade (Lamiaceae). *Conservation Science and Practice* 4(3): e621.

Nelson, G., P. Sweeney, and E. Gilbert. 2018. Use of globally unique identifiers (GUIDs) to link herbarium specimen records to physical specimens. *Applications in Plant Sciences* 6(2): e1027.

Owens, H., V. Barve, and S. Chamberlain. 2023. spocc: Interface to species occurrence data sources. Website: https://CRAN.R-project.org/package=spocc [accessed 13 February 2024].

Page, L. M., B. J. MacFadden, J. A. Fortes, P. S. Soltis, and G. Riccardi. 2015. Digitization of biodiversity collections reveals biggest data on biodiversity. *BioScience* 65(9): 841–842.

Panter, C. T., R. L. Clegg, J. Moat, S. P. Bachman, B. B. Klitgård, and R. L. White. 2020. To clean or not to clean: Cleaning open–source data improves extinction risk assessments for threatened plant species. *Conservation Science and Practice* 2(12): e311.

Park, D. S., G. M. Lyra, A. M. Ellison, R. K. B. Maruyama, D. dos Reis Torquato, R. C. Asprino, B. I. Cook, and C. C. Davis. 2023. Herbarium records provide reliable phenology estimates in the understudied tropics. *Journal of Ecology* 111(2): 327–337.

Pearson, K. D., G. Nelson, M. F. J. Aronson, P. Bonnet, L. Brenskelle, C. C. Davis, E. G. Denny, et al. 2020. Machine learning using digitized herbarium specimens to advance phenological research. *BioScience* 70(7): 610–620.

Phillips, S. J. 2017. A brief tutorial on Maxent. Website: https://biodiversityinformatics.amnh.org/open_source/maxent/Maxent_tutorial2017.pdf [accessed 13 February 2024].

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling* 190(3–4): 231–259.

Rawal, D. S., S. Kasel, M. R. Keatley, and C. R. Nitschke. 2015. Herbarium records identify sensitivity of flowering phenology of eucalypts to climate: Implications for species response to climate change. *Austral Ecology* 40(2): 117–125.

Ribeiro, B., S. Velazco, K. Guidoni-Martins, G. Tessarolo, and L. Jardim. 2023. bdc: Biodiversity Data Cleaning. Website: https://CRAN.R-project.org/package=bdc [accessed 13 February 2024].

Ritter, C. D., A. Zizka, C. Barnes, R. H. Nilsson, F. Roger, and A. Antonelli. 2019. Locality or habitat? Exploring predictors of biodiversity in Amazonia. *Ecography* 42(2): 225–399.

Soltis, D. E., and P. S. Soltis. 2016. Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. *Plant Diversity* 38(6): 264–270.

Soltis, P. S. 2017. Digitization of herbaria enables novel research. *American Journal of Botany* 104(9): 1281–1284.

Steen, V. A., M. W. Tingley, P. W. C. Paton, and C. S. Elphick. 2021. Spatial thinning and class balancing: Key choices lead to variation in the performance of species distribution models with citizen science data. *Methods in Ecology and Evolution* 12(2): 213–390.

Su, J.-X., C.-C. Dong, Y.-T. Niu, L.-M. Lu, C. Xu, B. Liu, S.-L. Zhou, et al. 2020. Molecular phylogeny and species delimitation of Stachyuraceae: Advocating a herbarium specimen-based phylogenomic approach in resolving species boundaries. *Journal of Systematics and Evolution* 58(5): 710–724.

van Proosdij, A. S. J., M. S. M. Sosef, J. J. Wieringa, and N. Raes. 2016. Minimum required number of specimen records to develop accurate species distribution models. *Ecography* 39(6): 542–552.

Veloz, S. D. 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography* 36(12): 2290–2299.

Wang, A., A. E. Melton, D. E. Soltis, and P. S. Soltis. 2022. Potential distributional shifts in North America of allelopathic invasive plant species under climate change models. *Plant Diversity* 44(1): 11–19.

Westgate, M., D. Kellie, M. Stevenson, and P. Newman. 2023. galah: Biodiversity data from the Living Atlas Community. Website: https://CRAN.R-project.org/package=galah [accessed 13 February 2024].

White, E., P. S. Soltis, D. E. Soltis, and R. Guralnick. 2023. Quantifying error in occurrence data: Comparing the data quality of iNaturalist and digitized herbarium specimen data in flowering plant families of the southeastern United States. *PLoS ONE* 18(12): p.e0295298.

Wickham, H., R. François, L. Henry, K. Müller, and D. Vaughan. 2023. Dplyr: A grammar of data manipulation. Website: https://CRAN.R-project.org/package=dplyr [accessed 13 February 2024].

Wieczorek, J. R., Q. Guo, and R. Hijmans. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science* 18(8): 745–767.

Wieczorek, J., D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, and D. Vieglais. 2012. Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE* 7(1): e29715.

Willis, C. G., E. R. Ellwood, R. B. Primack, C. C. Davis, K. D. Pearson, A. S. Gallinat, J. M. Yost, et al. 2017. Old plants, new tricks: Phenological research using herbarium specimens. *Trends in Ecology & Evolution* 32(7): 531–546.

Wollan, A. K., V. Bakkestuen, H. Kauserud, G. Gulden, and R. Halvorsen. 2008. Modelling and predicting fungal distribution patterns using herbarium data. *Journal of Biogeography* 35(12): 2298–2310.

Wu, Y., A. L. Hipp, G. Fargo, N. Stith, and R. E. Ricklefs. 2023. Improving species delimitation for effective conservation: A case study in the endemic maple-leaf oak (*Quercus acerifolia*). *New Phytologist* 238(3): 1278–1293.

Yao, X. A. 2020. Georeferencing and geocoding. In: A. Kobayashi [ed.], *International Encyclopedia of Human Geography*, 2nd ed. 111–117. Elsevier, Oxford, United Kingdom.

Zapata, F., and I. Jiménez. 2012. Species delimitation: Inferring gaps in morphology across geography. *Systematic Biology* 61(2): 179–194.

Zizka, A., D. Silvestro, T. Andermann, J. Azevedo, C. Duarte Ritter, D. Edler, H. Farooq, et al. 2019. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution* 10(5): 744–751.

Zizka, A., F. A. Carvalho, A. Calvente, M. R. Baez-Lizarazo, A. Cabral, J. F. R. Coelho, M. Colli-Silva, et al. 2020. No one-size-fits-all solution to clean GBIF. *PeerJ* 8: e9916.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Complete gatoRs workflow demonstrated in an R script.

Appendix S2. Complete gatoRs workflow demonstrated in an R markdown.

Appendix S3. Overview of functions with information on purpose, input/output, and dependencies.

Appendix S4. Number of records downloaded in total and from each aggregator (GBIF and iDigBio) for 25 plant species when using gatoRs, spocc with a download limit of 500, and spocc with a download limit of 100,000.

Appendix S5. Number of records in total and from each aggregator (GBIF and iDigBio) after within-aggregator duplicates were removed for 25 plant species when using gatoRs, spocc with a download limit of 500, and spocc with a download limit of 100,000.

Appendix S6. Number of records retained for 25 plant species through taxonomic filter, locality filter, and duplicate removal for records downloaded using gatoRs, spocc with a download limit of 500, and spocc with a download limit of 100,000.

Appendix S7. After records downloaded using gatoRs and spocc with a download limit of 100,000 were processed (taxonomic filter, locality filter, duplicate removal, removal of missing locality, and removal of occurrence records attributed to spatial error), we defined a convex hull that intersected with global land mass for each data set. Convex hulls representing gatoRs records are shaded orange, while convex hulls representing spocc are shaded blue.

How to cite this article: Patten, N. N., M. L. Gaynor, D. E. Soltis, and P. S. Soltis. 2024. Geographic And Taxonomic Occurrence R-based Scrubbing (gatoRs): An R package and workflow for processing biodiversity data. *Applications in Plant Sciences* 12(2): e11575. https://doi.org/10.1002/aps3.11575