LC-ABQ-Biopark Script Instructions

Introduction

The LC-ABQ-Biopark Script is intended to quickly screen North American plant species that are likely to be considered Least Concern and quickly compile basic data needed for an assessment. The tool creates .csv tables for upload using SIS Connect. At present, the tool is designed for use with North American plant species, but it has a modular framework, and can be modified to incorporate data from other sources. The tool is currently undergoing beta testing and is likely to have numerous errors. Should you encounter errors, please report them to Clay Meredith ([clay@bioparksociety.org](mailto:clay@bioparksociety.org)). The script has been extensively tested on computers running Windows, but is only periodically run on Apple products. Running the script on Apple products may produce unpredictable results.

Required Software

The script runs using the RStudio package. This requires installation of R (available free at: <https://www.r-project.org/>) and RStudio (available free at: <https://posit.co/download/rstudio-desktop/>). The main package (LC\_pipeline\_main.R) can be run in the RStudio interface.

User Inputs

User inputs can be modified within the User\_Inputs folder. There is one required user input and one set of optional inputs. A species list is required for the script to run. Users can input this value in one of two formats. The simplest is to download the taxonomy.csv file associated with an SIS working set. This file must be named taxonomy.csv and should be added to the User\_Inputs folder. Alternatively, a user can enter their own list using the spec\_list\_template.csv. This template file contains the appropriate headers for the script. The iucn\_id field is optional, but can be prepopulated with the id number for the species obtained from SIS. The Species field should contain individual binomial species names. The authority for the species is optional and should be included in the author field if it is to be used.

Once data is entered, save this file as spec\_list.csv in the User\_Inputs folder. If the script encounters both a taxonomy.csv file and a spec\_list.csv file, it will defer to the taxonomy.csv file and ignore the spec\_list.csv file. Since the number of GBIF records downloaded cannot be throttled, input of very common species may consume large amounts of system resources. For this reason, it is advisable to run batches of around 20 species at a time to avoid crashing the program. Users are given the option to subset the species list within the script if the number of species in the input file is large.

DarwinCore data files (used in lieu of or in addition to GBIF occurrence data) are an optional user input. These can be downloaded from a variety of museums and repositories and can be saved in the folder User\_Inputs\DarwinCore\_files. Outputs from Symbiota Network data sources can be entered directly as .zip files, or unpacked to add individual occurrences.csv files. The script will disregard all naming conventions associated with files in the User\_Inputs\DarwinCore\_files folder, but will not search subfolders within this directory. The species associated with each record will be read from the DarwinCore file structure. Please note that any records listed under a synonym may not be included in the analysis.

Script Outputs

Output .csv files are saved by the script to the LC-ABQ-Biopark\Outputs directory. Output files are formatted for upload to SIS using SIS Connect. The data files follow a standard naming convention with a unique batch number appended to the end of the file name. Each run of the script will be assigned a batch number and this can be used to identify different components of the output. A tool is available to automatically collate multiple batches into a single set of files for SIS Connect upload (see Batch Collate below). Please note that the script generates a comprehensive list of synonyms that is not appropriate for upload to SIS. This file is provided for the user’s reference but should be omitted from SIS Connect uploads.

In addition to the standard SIS Connect files, several other reference files are generated. The first of these is labeled DC\_raw\_data and is a reference file allowing the user to examine which DarwinCore records were utilized in the analysis. A parameters file is also exported. This contains data on which choices the user made in the run for an individual batch. This is useful for troubleshooting and for record keeping. Point\_data files include reformatted point data derived from DarwinCore and/or GBIF locality data. This file is formatted according to the Red List mapping standards and can be submitted as a map for the species. Finally, the species\_inputs file documents which species were submitted by the user in an individual batch. This is augmented with data pertinent to the species including estimations of Area of Occupancy, Extent of Occurrence, indicators of whether external DarwinCore data was included in the analysis, and a record of taxonomic comparisons made by the script. For example, the script searches for taxonomic comparisons against Flora of North America, IT IS, VASCAN, and Kew’s Plants of the World Online. If searches for input species produce exact matches, this is noted. If the species is instead listed as a synonym by the data source, the synonym used is indicated. Lack of taxonomic match may also be indicated here.

Running the Script

The package is controlled by the LC\_pipeline\_main.R file. Opening the script in R Studio will produce a screen that looks like the one below (note that there may be some subtle differences, based on how R Studio is configured on your machine).

A screenshot of a computer

Description automatically generated

The software is configured to run based on the file structure in which it was downloaded. The main folder can be placed anywhere on your machine, but the underlying file structure should not be altered beyond the prescribed changes to the User\_Inputs folder. Once a data source has been supplied, the script can be run by pressing the “Source” button (circled in the above image).

Pressing the source button will initiate several processes which will be visible in the Console window. The first time the script is run, it will download some geospatial data used in calculations, and packages used for computation within RStudio. This may take a few minutes. Subsequent runs of the software will use the previously downloaded versions and will run much faster. Once base data has been downloaded, the script will ask the user to select whether GBIF data should be used in the analysis. If the user selects no, only DarwinCore data supplied by the user will be used for spatial analysis.

If the user selects yes, the script will ask further questions to access your GBIF account. The script will then prompt the user to enter their GBIF user name, email address, and password. The password is saved in an encrypted format on your computer. It can be stored over multiple uses and is only removed upon running the Downloaded Data Clear script. This must be run to switch between GBIF user accounts.

Once GBIF information has been entered, the script will prompt the user to enter a compiler name. This name will be entered into the compiler field in the point data for upload as an accompanying map for the species. The user will then be prompted to choose if iNaturalist records should be used in the analysis. Research grade occurrences are downloaded with GBIF records, but may not be suitable for all analyses. Finally, the user is prompted to subset the data. Entering a long list of species may crash R. It is recommended that batches of 20 species be run together. This will prevent species with many observations from overwhelming the spatial analysis. If the user chooses to subset data, they will be prompted to enter a minimum and maximum row number for the analysis.

The analysis will then run. Downloading GBIF data may take some time, and the analysis will run as much as possible in the meantime. Users may be left with

Data Sources

The tool compiles assessment data from various online platforms including the Integrated Taxonomic Information System (ITIS), Global Biodiversity Information Facility (GBIF.org), Kew’s Plants of the World Online, NatureServe Explorer, the Database of Vascular Plants of Canada (VASCAN), and Flora of North America (FNA). Additional modules are under development and may be available in future iterations of the script. User supplied DarwinCore files may also be used to augment these sources.

Batch Collate

Files from multiple batches may be collated using the batch\_collate.R script. This will produce a collated set of files in the Outputs folder. The collate process will not impact existing batch files, but will overwrite already collated batches. The script requires no user input and can be run using the “Source” button at the top of the RStudio console window.

Downloaded Data Clear

The script stores significant quantities of data on the host machine. Users can clear this data (along with stored GBIF password, and downloaded geospatial datasets) by running the Downloaded\_data\_clear.R script. This script requires no user input and can be run using the “Source” button at the top of the RStudio console window.