丁刘勇, 李昊, 陶捐, 张金龙, 黄敏睿, 杨科, 王军, 丁城志, 何大明 (2021) 获取生物物种名录信息的 R 程序包 SP2000. 生物多样性, 29 (1): 118 – 122. http://www.biodiversity-science.net/CN/10.17520/biods.2020235

SP2000: An open-sourced R package for querying the Catalogue of Life

Liuyong Ding[1,2], Hao Li[3], Juan Tao[1,2], Jinlong Zhang[4], Minrui Huang[1,2], Ke Yang[1,2,4], Jun Wang[1,2], Chengzhi Ding[1,2*], Daming He[1,2*]

1 *Institute of International Rivers and Eco-security, Yunnan University, Kunming 650504*
2 *Yunnan Key Laboratory of International Rivers and Transboundary Eco-security, Yunnan University, Kunming 650504*
3 *National Pilot School of Software, Yunnan University, Kunming 650504*
4 *Flora Conservation Department, Kadoorie Farm and Botanic Garden, Hong Kong 999077*

ABSTRACT

Aims: The Catalogue of Life provides the basis for understanding both regional and global biodiversity. With the invention and development of the internet, the up-to-date species checklists stored in the public databases has greatly promoted the development of taxonomy, conservation biology, and macroecology. Public species checklists play an indispensable role in biodiversity conservation and aid in the assessment of species' conversation status. The Species 2000 China Node (http://www.sp2000.org.cn) and the Catalogue of Life (http://www.catalogueoflife.org) are among the leading online databases in cataloguing biodiversity, contain 122,280 and 1,829,672 taxa respectively (including infraspecific taxa). Although searching the content of the websites may be relatively straightforward, downloading the data and transferring it into a statistical environment for further analysis can present challenges.

Method: To address this issue, we developed the package SP2000 using the R programming language.

Application: SP2000 is an open-source, cross-platform, and user-friendly package which aims to help users query and download the checklist of organisms (including animals, plants, fungi, and microbes) from within and outside China. Here we introduce and describe the usage of SP2000 including installation, and configuration of parameters.

Key words: species checklist; redlist; China's biodiversity; R package

1 Introduction

The Catalogue of Life provides the basis for understanding both regional and global biodiversity (Reichhardt, 1999; Banki et al, 2019; Ower & Roskov, 2019). With the invention and development of the internet, Annual or monthly editions of species checklists have been numerously stored in the public databases (e.g. the Species 2000 China Node and Catalogue of Life), which have greatly promoted the development of taxonomy, conservation biology and macroecology (Jiang et al, 2015). At present, these databases have been widely used in the assessment of species status, red list compilation and biodiversity conservation for governments or international organizations.

What are the similarities and differences between the Species 2000 China Node and Catalogue of Life? Their goals are to provide a validated checklist of the known species to all users in the world. As of June 4, 2020, the Species 2000 China Node (http://www.sp2000.org.cn, the Biodiversity Committee of Chinese Academy of Sciences, 2020) and the Catalogue of Life (http://www.catalogueoflife.org) record 122,280 and 1,829,672 taxa (including infraspecific taxa), respectively. The latter is a collection of more than 130 global species databases, while the former is a subset of national studies. Therefore, the spatial scale of the both checklists is different and complementary. The Species 2000 China Node was established in 2006 by the Biodiversity Committee of the Chinese Academy of Sciences (BCCAS), which is contributed by the Institute of Botany, Institute of Microbiology, Institute of Oceanology and Institute of Zoology, CAS. The first annual checklist of Catalogue of Life China was released in 2008 and has been updated annually, which is an important data source for the Catalogue of Life annual checklist (Jiang et al, 2015; Ma et al, 2018). Compared with global Catalogue of Life, Catalogue of Life China also provides the Chinese name (i.e., characters and pinyin) in addition to containing the scientific name of each species, synonyms, alias, references, classification system, distribution area and other information. Although all information in two websites will be available to all users in the world freely, downloading and getting required data into a statistical environment for further analysis are not straight-forward, which has become the main obstacle to restrict the widespread use of these checklists.

To address the above problems, we developed the R package SP2000 using R programming language (due to its features of open-sourced, cross-platformed,

etc.; Tippmann, 2014; Zhang et al, 2016; Lai et al, 2019; R Core Team, 2020), which aims to help users accurately, quickly to query and download the required species checklists from the Species 2000 China node and Catalogue of Life website.

2 Methods

The R package *SP2000*, a programmatic interface to http://sp2000.org.cn, re-written based on an accompanying 'Species 2000' API (http://sp2000.org.cn/api/document, version 2), and access tables describing catalogue of the Chinese known species of animals, plants, fungi, micro-organisms, and more. This package also supports access to Catalogue of Life (http://webservice.catalogueoflife.org/col/webservice, version 1.9).

Compared to other tools for acquiring species checklists such as Catalogue of Life Search Plugin (http://www.catalogueoflife.org/content/web-browser-page-plugin), CD-ROM of Catalogue of Life China (http://sp2000.org.cn/download) and portal-components developed via JavaScript language (https://github.com/CatalogueOfLife/portal-components), the SP2000 is an open-sourced, cross-platformed, and user-friendly package which aims to help users to query and download the checklist of animals, plants, fungi and micro-organisms both in and outside China. The downloaded information goes directly into R statistical environment for further analysis, for example, mining more information of biodiversity through R package spocc (Chamberlain, 2020).

3 Usage

3.1 Version and installation

The package *SP2000* written in R language has been submitted to the CRAN (https://cran.r-project.org/package=SP2000, version 0.1.0), users can easily install R packages SP2000 using R commands install.packages("SP2000", repos = "https://cran.r-project.org"). It mainly consists of eight functions covering set_search_key, search_family_id, search_taxon_id, search_checklist, get_redlist_China, get_col_global, find_synonyms and get_col_taiwan, and the configuration of the parameters for theses function are as follows.

3.2 set_search_key

This function set_search_key allows users to set the key variable used all search_* functions (e.g. search_family_id, search_taxon_id and search_checklist). Users can obtain a key by registering at http://sp2000.org.cn/api/document, clicking on the user information. It is worth noting that the upper limit of daily API visits for ordinary users is 2000, and users can apply for increasing the daily API request limit, through filling in the application form http://col.especies.cn/doc/API.docx and send an email to SP2000CN@ibcas.ac.cn entitled "Application for increasing API Request Times". Set the API key by running the R code:

```
set_search_key <- "your apikey"
```

3.3 search_family_id

The family is the most commonly used classification rank in biological classification, through which the taxonomic unit of species or subspecies can be more easily inquired. The Species 2000 China Node defines unique idenity (id) for family and species (subspecies) to ensure the accuracy of data query. The function search_family_id provides the function querying the collection of family's ids. There are four arguments including query, start, limit and mc.cores. (1) The parameter 'query' supports one or more queries for family name, or part of family name, (2) the parameter 'start' sets the number of record to start at, the default value of 1, (3) the parameter 'limit' sets the number of records to return, the default value is 20, and (4) the parameter 'mc.cores' can set the number of cores to use, the default value is 2. Search family ids by running:

```
search_family_id (query = "Anguillidae")
```

3.4 search_taxon_id

The family ids can be used to directly obtain the list of ids for the species or subspecies, and then the details of the species list can be obtained using the function search_checklist. The search_taxon_id supports multiple types of queries for family's ids, scientific name and common name (including Chinese name). There are five arguments including query, name, start, limit and mc.cores. (1) The parameter query supports one or more queries, (2) the parameter name sets the query mode, in conjunction with the parameter query, parameters to be selected are "familyID","scientificName" and "commonName", the default value is "scientificName", (3) the parameter start sets the number of record to start at, the default value of 1, (4) the parameter limit sets the number of records to return, the default value is 20, and (5) the parameter mc.cores setting is the same as 3.3. Take "Anguillidae" as an example, the R code is as follows:

```
## loading package
library("SP2000")
## Set your Species 2000 API key
set_search_key <- "your apikey"
## Search family ids via family name
```

familyid <- search_family_id (query = "Anguillidae")

Search taxon ids via family's ids

query <- familyid$Anguillidae$data$record_id

taxonid <- search_taxon_id (query = query, name = "familyID")

3.5 search_checklist

The function search_checklist gets detailed information of species through species ids, including scientific name, synonym, alias, literature, classification system, distribution region and other data, as well as Chinese name and Chinese name pinyin and other contents. This function needs to be used in combination with the functions search_family_id and search_taxon_id. There are two arguments including query and mc.cores. (1) The parameter query supports one or more queries and (2) The parameters mc.cores is the same as 3.3.

Take the query result of 3.4 as an example：

query <- taxonid[["3851c5311bed46c19529cb155d37aa9b"]][["data"]][["namecode"]]

search_checklist (query = query)

3.6 get_redlist_china

The function get_redlist_china has four parameters: query, option, group, and viewDT. (1) The parameter 'query' supports one or more queries for scientific name or Chinese name, (2) the parameter 'option' sets the query mode, which is used in conjunction with the parameter query. The parameters option include "Chinese Names" and "Scientific Names", and the default value is "Scientific Names", (3) the optional parameters 'group' includes "Amphibians," "Birds", "Mammals", "Inland Fishes", "Reptiles," " Plants" and "Fungi", and (4) the parameter 'viewDT' is the logical value, which is used together with the parameter group. If viewDT = TRUE, the query result will display an interactive page. Taking Inland Fishes as an example, it is called get_redlist_China (... , group = "Inland Fishes", viewDT = TRUE).

Take the *Anguilla* query as an example, the R code is as follows:

Get Chinese Red List of the genus *Anguilla*

get_redlist_china (query = "Anguilla", option = "Scientific Names")

Query "Inland Fish" China Red List information displaying searchable, downloadable and interactive page

get_redlist_china (group = "Inland Fishes", viewDT = TRUE)

3.7 get_col_global

The function get_col_global is unrestricted by the Species 2000 key and can be used independently. It contains six parameters: query, option, response, start, limit and mc.cores. (1) The parameter 'query' inputs one or more ids or the species name, (2) the parameter 'option' sets the query mode, which is used in conjunction with the parameter query. The optional parameters have "ID" and "name", and the default value is "name", (3) the parameter 'response' sets the query to return the result, which can be selected as "Full", one of "terse". "Full" returns the full query result, and "terse" returns the short query result. The default value is "terse"; (4) The parameter 'start' sets the first record returned by the query. The default value is 0, which is used in conjunction with the parameter 'response', (5) the parameter 'limit' sets the record returned by a single query, the default value is 500, the maximum number of results returned by a single short query is 500, and the maximum number of results returned by a single complete query is 50, and (6) the parameter 'mc.cores' setting is the same as 3.3.

Take the *Anguilla* query as an example, the R code is as follows:

x <- get_col_global (query = "Anguilla", response = "full")

The total query result is 208

x[["Anguilla"]][["meta"]][["total_number_of_results"]] [1]

3.8 find_synonyms

The function find_synonyms has two arguments query and mc.cores. (1) The argument 'query' enters one or more species name, and (2) the argument 'mc.cores' is the same as 3.3. Take "*Anguilla Anguilla* "as an example and call it find_synonyms ("Anguilla Anguilla ").

3.9 get_col_taiwan

The function get_col_taiwan has four parameters: query, level, option and include_synonyms. (1) The parameter 'query' supports one or more queries, (2) the parameter 'level', which can be used in combination with the parameter query to select one of "kingdom","phylum","class","order","family","genus" and "species", (3) the parameter 'option' includes "Contain", "Equal" and "beginning", the default is "equal", and (4) the parameter 'include_synonyms' is the logical value, and the query result contains synonym information, with the default value of TRUE. Take Anguillidae as an example, the call method is

get_col_taiwan (query = "Anguillidae", level = "family").

4 Conclusion

In this paper, we presented a new tool (R package *SP2000*) of re-written via Web API, which provides an interface for application program to download the data of the species checklists. Its detailed usage makes

SP2000 a useful tool for biodiversity researchers and taxonomists.

In addition to R package *SP2000*, we also developed Python package *SP2000* (https://pypi.org/project/SP2000) through Python programming language (Perkel, 2015; Python Software Foundation, 2020) to better meet the needs of users in the era of big biodiversity data (Bisby, 2000). Users can easily install python packages SP2000 using commands "install pip3 install SP2000" or "python3 -m pip install SP2000". The configuration of the parameters and query are basically the same as that of R package *SP2000*. Future work on *SP2000* includes enhancements such as search for more taxonomic information of insects and invertebrates distributed in China. We will also add China Animal Scientific Database (http://zoology.especies.cn) to the package *SP2000* via Web API (http://zoology.especies.cn/database/api).

Acknowledgements

The R package *SP2000* was made possible by leveraging integral R packages including *jsonlite* (Ooms, 2014), *tibble* (Müller & Wickham, 2020), *rlist* (Ren, 2016) and many others. We thank two anonymous reviewers whose helpful feedback helped improve the package and clarify this manuscript.

Availability

The R and Python package SP2000 are freely available under the permissive MIT license at https://cran.r-project.org/package=SP2000 and https://pypi.org/project/SP2000, respectively.

ORCID

Liuyong Ding https://orcid.org/0000-0002-5490-182X
Jinlong Zhang https://orcid.org/0000-0002-1161-5460
Jun Wang https://orcid.org/0000-0003-2481-1409
Chengzhi Ding https://orcid.org/0000-0001-5215-7374

References

Banki O, Hobern D, Döring M, Remsen D (2019) Catalogue of Life Plus: A collaborative project to complete the checklist of the world's species. Biodiversity Information Science and Standards, 3, e37652.

Bisby FA (2000) The quiet revolution: Biodiversity informatics and the internet. Science, 289, 2309–2312.

Chamberlain S (2020) spocc: Interface to Species Occurrence Data Sources. R package version 1.0.8. https://CRAN.R-project.org/package=spocc/. (accessed on 2020-06-01)

Jiang ZG, Qin HN, Liu YN, Ji LQ, Ma KP (2015) Protecting biodiversity and promoting sustainable development: In memory of the releasing of Catalogue of Life China 2015 and China Biodiversity Red List on the International Day for Biological Diversity 2015. Biodiversity Science, 23, 433–434. (in Chinese)

Lai JS, Lortie CJ, Muenchen RA, Yang J, Ma KP (2019) Evaluating the popularity of R in ecology. Ecosphere, 10, e02567.

Ma KP, Zhu M, Ji LQ, Ma JC, Guo QH, Ouyang ZY, Zhu L (2018) Establishing China Infrastructure for Big Biodiversity Data. Bulletin of the Chinese Academy of Sciences, 33(8), 80–87. (in Chinese with English abstract)

Müller K, Wickham H (2020). tibble: Simple Data Frames. R package version 3.0.3. https://CRAN.R-project.org/package=tibble. (accessed on 2020-08-01)

Ooms J (2014) The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv, 1403, 2805.

Ower G, Roskov Y (2019) The Catalogue of Life: Assembling data into a global taxonomic checklist. Biodiversity Information Science and Standards, 3, e37221.

Perkel JM (2015) Programming: Pick up Python. Nature, 518, 125–126.

Python Software Foundation (2020) Python Language Reference, version 3.7. https://www.python.org/. (accessed on 2020-06-01)

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/. (accessed on 2020-06-01)

Reichhardt T (1999) Catalogue of life could become reality. Nature, 399, 519-519.

Ren K (2016) rlist: A Toolbox for Non-Tabular Data Manipulation. R package version 0.4.6.1. https://CRAN.R-project.org/package=rlist. (accessed on 2020-08-01)

The Biodiversity Committee of Chinese Academy of Sciences (2020) Catalogue of Life China: 2020 Annual Checklist, Beijing, China (in Chinese and in English). http://www.sp2000.org.cn/CoLChina. (accessed on 2020-05-22)

Tippmann S (2014) Programming tools: Adventures with RNature, 517(7532), 109–110.

Zhang JL, Zhu HL, Liu JG, Fischer GA (2016) Principles behind designing herbarium specimen labels and the R package 'herblabel'. Biodiversity Science, 24, 1345–1352. (in Chinese with English abstract)