

## pacote **Rocc**: gestão e análise de dados de ocorrências de espécies

O pacote **Rocc**(<https://liibre.github.io/Rocc/index.html>) disponibiliza ferramentas para criar fluxos de trabalho reprodutíveis para tratar dados abertos de ocorrência de espécies. Atualmente existem diversas iniciativas de compilação de bases de dados digitais de biodiversidade, como GBIF no âmbito global e speciesLink no Brasil. Um grande volume de dados está disponível de forma aberta na rede, porém, converter a informação disponível em um dado a ser utilizado em um estudo exige uma combinação de rotinas automatizadas e conhecimento contextual para tomada de decisões. Este pacote se propõe a auxiliar a geração de rotinas automatizadas de obtenção e checagem de dados de biodiversidade usando ferramentas como speciesLink (<https://specieslink.net/>), Flora do Brasil (<http://floradobrasil.jbrj.gov.br/reflora/listaBrasil>) e GADM (<https://gadm.org/data.html>). O objetivo do pacote é automatizar tarefas repetidas e facilitar a tomada de decisão por parte de especialistas e não exclui a necessidade de uma avaliação contextual dos dados. O pacote permite estabelecer um fluxo de trabalho desde a obtenção do dado até a checagem de nomenclatura taxonômica e geográfica e também permite que cada uma das funções seja utilizada de forma independente.

O pacote **Rocc** possui três grandes eixos: obtenção de dados de ocorrência, checagem taxonômica e obtenção de dados geográficos. No eixo de obtenção de dados, o pacote possui uma função para fazer download de dados da base do speciesLink, `rspecieslink()`. Essa função se conecta com a API do speciesLink (<https://api.splink.org.br/>) e permite fazer o download de dados a partir de nomes científicos de espécies ou de municípios brasileiros e na América Latina. Na página do pacote, há um artigo com exemplos de diferentes usos da função *Downloading speciesLink data from R* ([https://liibre.github.io/Rocc/articles/articles/using\\_rspecieslink.html](https://liibre.github.io/Rocc/articles/articles/using_rspecieslink.html)). O pacote possui também a função `rgbif2()`, uma função para fazer o download de dados de ocorrência do GBIF (<https://www.gbif.org/>), que encapsula uma função do pacote **rgbif** (Chamberlain et al 2021) para criar uma saída compatível com o fluxo de trabalho. Uma vez tendo o download de dados das duas fontes, é possível combiná-los usando a função `bind_dwc()`. Isso permite à pessoa usuária obter e combinar dados de diferentes fontes que vêm de origem seguindo padrões diferentes. Os dados são formatados em uma tabela seguindo os campos do formato DarwinCore, o formato padrão para dados de biodiversidade.

Além disso, o pacote permite comunicação direta com a base de dados do projeto Flora do Brasil (Flora do Brasil 2020), que é uma iniciativa de catalogação das plantas do Brasil desenvolvida por um comitê de especialistas da área de botânica. **Rocc** possui uma função para fazer o download da base de dados completa da Flora do Brasil, `update_flora()`. É possível também fazer a busca de listas de espécies seguindo critérios de domínio fitogeográfico onde a espécie ocorre, endemismo, forma de vida ou tipo de vegetação usando a função `search_flora()`. O artigo *Searching for species names in FB2020* ([https://liibre.github.io/Rocc/articles/articles/searching\\_flora.html](https://liibre.github.io/Rocc/articles/articles/searching_flora.html)) contém exemplos de como usar essas funções.

Como o dado de biodiversidade corresponde a uma espécie em um dado local, é bastante usual passar os dados por uma checagem geográfica. Como existem diversos pacotes de R desenvolvidos para esse fim, como o **coordinateCleaner** (Zizka et al 2017) por exemplo, o pacote **Rocc** se propõe a efetuar o download de shapefiles e gazetteers que irão auxiliar no processo de checagem geográfica. No eixo de obtenção de dados geográficos o pacote **Rocc** possui três funções: `getGADM()`, `getGAZ()` e `getWDPA()`. A função `getGADM()` faz o download de shapefiles de unidades administrativas para países do globo a partir de GADM, <https://gadm.org/data.html> e retorna objetos de classe sf ou sp. A função `getGAZ()` faz o download do gazetteer do DIVA-GIS (<http://www.diva-gis.org/gData>). A função `getWDPA()` faz download de shapefiles da base de dados de áreas protegidas da IUCN para cada país (<https://www.protectedplanet.net>). Diferentemente das funções do pacote que se apoiam em bases como speciesLink e Flora do Brasil, esse eixo do pacote tem escopo global e pode ser aplicado em diferentes contextos. O artigo *Geographic data download with Rocc* ([https://liibre.github.io/Rocc/articles/articles/data\\_download.html](https://liibre.github.io/Rocc/articles/articles/data_download.html)) contém exemplos de uso das funções desse eixo.

A partir dos três eixos do pacote de obtenção de dados de biodiversidade, checagem taxonômica e obtenção de dados geográficos, é possível criar fluxos de trabalho reprodutíveis totalmente baseados em dados abertos para obtenção e limpeza de dados de biodiversidade. Na página do pacote existem artigos para guiar o desenvolvimento de fluxos de trabalho. Com isso, espera-se facilitar o acesso a diferentes bases e permitir autonomia às pessoas usuárias de criarem seus próprios fluxos de trabalho adaptados a sua necessidade. O pacote já vem sendo utilizado em projetos acadêmicos e consultorias de listagens de espécies regionais. Além disso, também tem sido utilizado em aulas de limpeza de dados de biodiversidade. Esperamos com isso, permitir que as pessoas usuárias possam focar seu trabalho mais em seu contexto e no significado biológico e menos em tarefas automatizadas de download e limpeza que já são feitas pelo pacote **Rocc**.

**Palavras-chave:** DIVA-GIS; flora do Brasil; GADM; GBIF; gazetteer; informática da biodiversidade; reprodutibilidade; speciesLink

## Referências

Chamberlain S, Barve V, Mcglinn D, Oldoni D, Desmet P, Geffert L, Ram K (2021). *rgbif: Interface to the Global Biodiversity Information Facility API*. R package version 3.5.0, <URL: <https://CRAN.R-project.org/package=rgbif>>.

Flora do Brasil 2020. Jardim Botânico do Rio de Janeiro. Available at: < <http://floradobrasil.jbrj.gov.br/> >. Accessed on: 28 Jul. 2021

Zizka A, Silvestro D, Andermann T, Azevedo J, Duarte Ritter C, Edler D, Farooq H, Herdean A, Ariza M, Scharn R, Svanteson S, Wengstrom N, Zizka V, Antonelli A (2019). "CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases." *Methods in Ecology and Evolution*, -7. doi: 10.1111/2041-210X.13152 (URL: <https://doi.org/10.1111/2041-210X.13152>), R package version 2.0-18, <URL: <https://github.com/ropensci/CoordinateCleaner>>.