AEET
ASOCIACIÓN ESPAÑOLA DE ECOLOGÍA TERRESTRE

Ecosistemas 26(1): 126-127 [Enero-abril 2017]
Doi.: 10.7818/ECOS.2017.26-1.20

NOTAS ECOINFORMÁTICAS

ecosistemas
REVISTA CIENTÍFICA DE ECOLOGÍA Y MEDIO AMBIENTE

ISSN 1697-2473 / Open access
disponible en www.revistaecosistemas.net

rOpenSci: cómo acceder de forma reproducible a repositorios de datos públicos

I. Bartomeus[1]*

(1) Departmento de Ecología Integrativa, Estación Biológica de Doñana, Consejo Superior de Investigaciones Científicas, Avda. Américo Vespucio 26, 41092 Sevilla, España.

* Autor de correspondencia: I. Bartomeus [nacho.bartomeus@gmail.com]

> Recibido el 30 de marzo de 2017 - Aceptado el 31 de marzo de 2017

Bartomeus, I. 2017. rOpenSci: cómo acceder de forma reproducible a repositorios de datos públicos. *Ecosistemas* 23(1): 126-127. Doi.: 10.7818/ECOS.2017.26-1.20

El uso de datos públicos ha abierto la puerta a responder un abanico de preguntas que antes era inimaginable (Reichman et al. 2011; Michener y Jones 2012; Hampton et al. 2013). En ecología hay repositorios de datos de ocurrencia de especies (www.gbif.com), climáticos (www.worldclim.com) o filogenéticos (www.treebase.com) que por sí solos o combinados permiten realizar análisis para testar teorías generales a una escala global. Sin embargo, obtener estos datos suele hacerse de forma no reproducible mediante el uso de interfaces web. Estas interfaces son útiles para explorar los datos y además permiten seleccionar qué tipo de datos quieres y descargarlos a tu ordenador, pero este proceso es subóptimo por dos razones. Primero, una vez tienes los datos, es imposible recordar qué selección hiciste exactamente (no hay trazabilidad; Rodríguez-Sanchez et al. 2016). Segundo, cuando hay nuevos datos disponibles o se han corregido errores en los datos originales del repositorio es imposible actualizar los datos que descargaste de forma automática.

El equipo de rOpenSci (https://ropensci.org), entre otras cosas, se dedica a crear paquetes de R que permiten el acceso a datos públicos desde tu terminal de R. Esto te permite tener un código legible tanto por máquinas como por humanos que especifica exactamente qué datos has descargado (hay una trazabilidad completa). Asimismo, actualizar los datos con los últimos cambios no requiere más esfuerzo que volver a correr el código. Las ventajas no están solo en incrementar la reproducibilidad de tus resultados, sino que sobre todo, el acceso a los datos directamente desde R te permite ahorrar tiempo y reducir el número de pasos de tu flujo de trabajo, y por consiguiente el número de potenciales errores de bulto que se pueden cometer en cada paso.

Un ejemplo de cómo usar estos paquetes es bajar datos de tu especie favorita de gbif usando el paquete rgbif. Vamos a verlo en tres simples pasos que apenas ocupan cuatro líneas de código en R.

1) Instalar y cargar el paquete.

```
install.packages("rgbif") #solo has de instalarlo la primera vez
library(rgbif)
```

2) Obtener los datos que queremos.

```
dat <- occ_search(scientificName = "Andrena flavipes", limit = 10000)
```

3) Mostrar gráficamente los datos obtenidos (**Fig. 1**).

```
library(mapr)
map_ggmap(dat, zoom = 4)
```

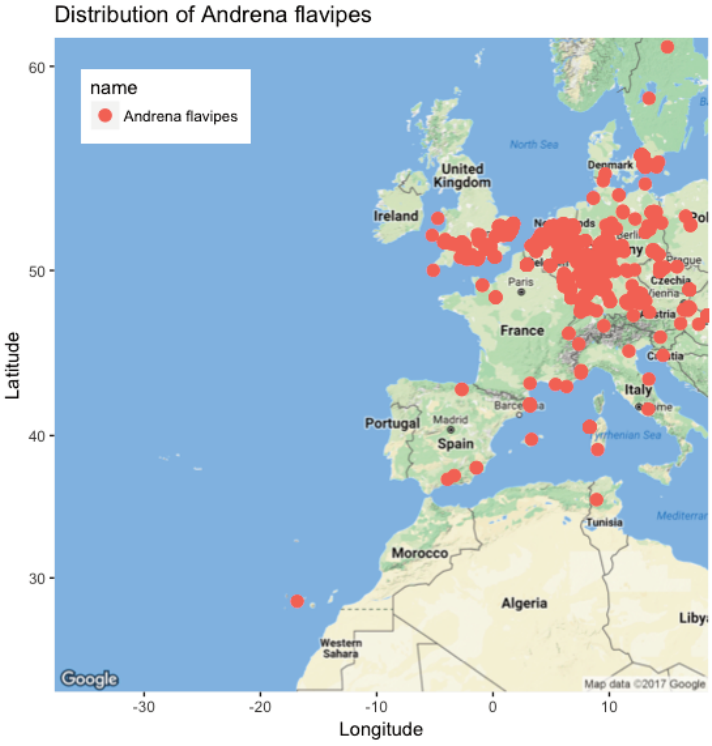

Figura 1. *Mapa de localidades de* Andrena flavipes *disponibles en GBIF, obtenido directamente desde R.*

En este ejemplo tan simple nos limitamos a representar datos gráficamente, pero se pueden hacer análisis más interesantes combinando datos de varias fuentes (e.j. https://ropensci.org/usecases/). Como advertencia, hay que tener en cuenta que este tipo de datos abiertos pueden contener información erronea que deberá ser depurada antes correr los análisis (Robertson et al. 2016). El paquete scrubr, en desarrollo, ayuda a realizar esta necesaria limpieza inicial para este tipo de datos (https://github.com/ropensci/scrubr).

La lista de paquetes disponibles para acceder a datos crece año a año (https://ropensci.org/packages/#data_access) y los paquetes se actualizan con regularidad solucionando errores o introduciendo mejoras. Por tanto, si encuentras algún fallo lo mejor que puedes hacer es comentárselo a sus autores.

Si estás interesado en el uso de datos libres y quieres hacerlo de forma programática, rOpenSci es un buen inicio, pero hay otros paquetes para R disponibles en CRAN, y para los que no usan R a veces existen clientes específicos para otros lenguajes (e.j. python). Solo es cuestión de saber buscar la herramienta que necesitas.

Agradecimientos

Han contribuido a su revisión: Carlos Lara, Ana Isabel García-Cervigón, Antonio Jesús Pérez Luque y Francisco Rodríguez-Sánchez.

Referencias

Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S., Porter, J.H. 2013. Big data and the future of ecology. *Frontiers in Ecology and the Environment* 11: 156-162.

Michener, W.K., Jones, M.B. 2012. Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology and Evolution* 27: 85-93.

Reichman, O.J., Jones, M.B., Schildhauer, M.P. 2011. Challenges and Opportunities of Open Data in Ecology. *Science* 331: 703-705.

Robertson, M.P., Visser, V., Hui, C. 2016. Biogeo: an R package for assessing and improving data quality of occurrence record datasets. *Ecography* 39: 394-401.

Rodríguez-Sanchez, F., Pérez-Luque, A.J., Bartomeus, I., Varela, S. 2016. Ciencia reproducible: qué, por qué, cómo? *Ecosistemas* 25: 83-92.