WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

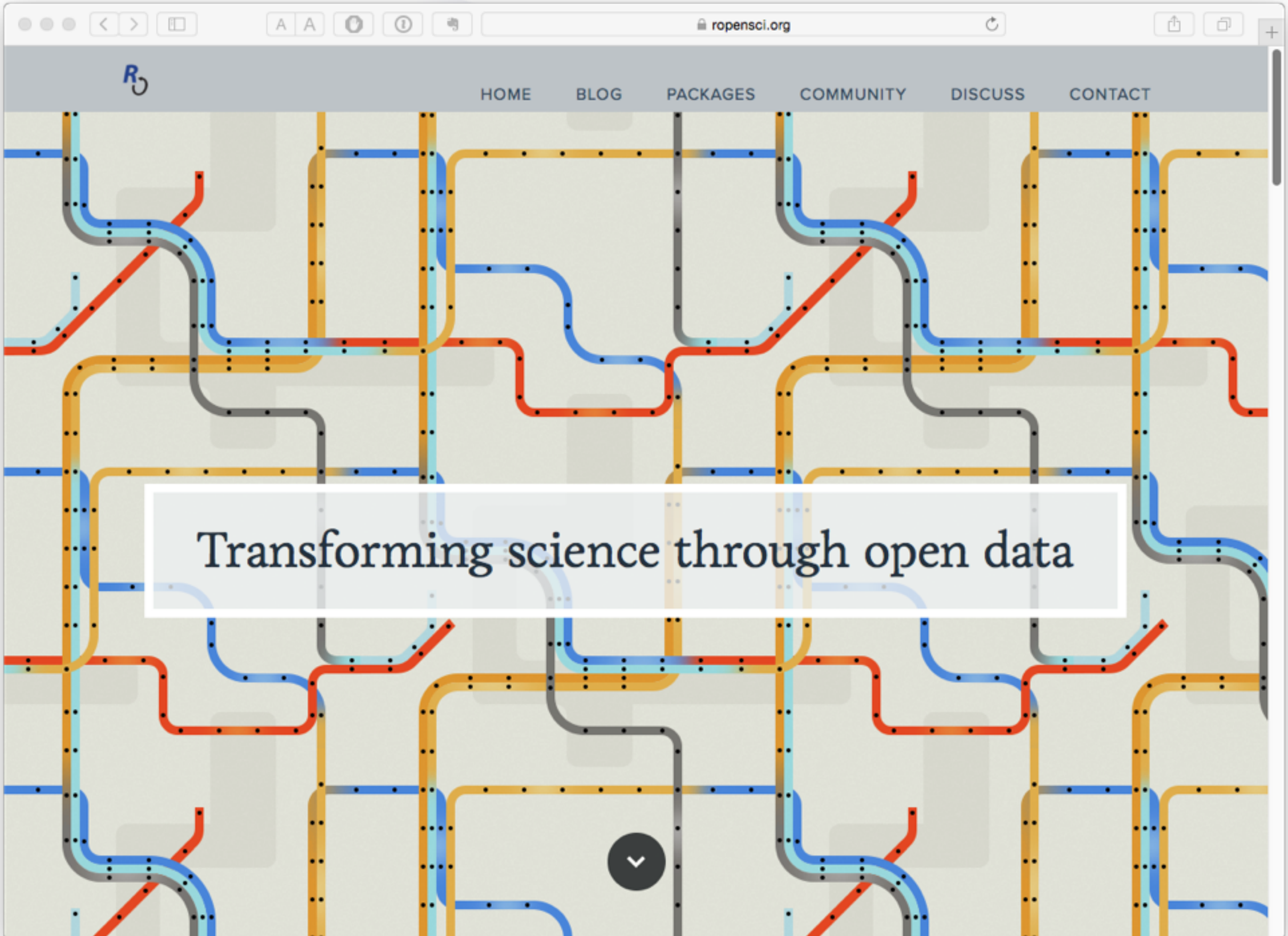
Accessing the Scientific Literature using APIs/web services

Paul Oldham

APIs (Application Programming Interfaces) or web services

- Application Programming Interfaces (APIs) sound a bit complicated:
  - Basically they are a URL or set of URLs that provide access to databases behind a web site using different routes (endpoints e.g. inventors, applicants).
  - The advantage of APIs is that it is normally possible to obtain much more data, of different types, than from a web page.
  - Challenges include understanding the endpoints, and translating the XML/JSON into data tables.
  - In RStudio a growing number of rOpenSci packages have already solved these problems for us.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION


ropensci.org
HOME
BLOG
PACKAGES
COMMUNITY
DISCUSS
CONTACT
Transforming science through open data

ropensci.org/packages/

HOME BLOG PACKAGES COMMUNITY DISCUSS CONTACT

rOpenSci packages

This is a complete list of all available rOpenSci packages. Packages are grouped by ones that acquire data, full-text of journal articles, altmetrics, data-publication, focus on scalable and reproducibile computing, data visualization, or geospatial work. Packages with a CRAN sign are stable versions that you can quickly install from your nearest mirror using `install.packages("PACKAGE_NAME")`. Others are in various stages of development (bleeding edge packages are not listed here) and you can learn more by following our GitHub organization page, and our GitHub organization for bleeding edge projects. All of our software packages are open source. Please see package description files for more information on specific licenses. We also have a package status dashboard that updates periodically.

Many of our packages are community-contributed. If you are interested in contributing a package, please visit our onboarding repository for details.

Data Publication | Data Access | Literature | Altmetrics | Scalable & Reproducible Computing | Databases | Data Vizualization | Data Tools | Geospatial

Leave us a message!

ropensci.org/packages/

Literature

Packages that interface with journal full-text or metadata

Package	Description	Details
aRxiv	Access to the aRxiv preprint repository	CRAN GITHUB
bmc	An R interface to Biomed Central (BMC) search API and full text XML. IN EARLY DEVELOPMENT	CRAN GITHUB
csl	Explore CSL styles and locales. IN EARLY DEVELOPMENT	CRAN GITHUB
fulltext	An high level R interface to many different sources for full text (or nearly) scholarly texts, including PLOS, PMC, PeerJ, eLife, arXiv, and many more.	CRAN GITHUB
IEEER	Interface to the IEEE Xplore Gateway. IN EARLY DEVELOPMENT	CRAN GITHUB
oai	Work with OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) services, a protocol developed by the Open Archives Initiative.	CRAN GITHUB
pdftools	A fast and portable package to extract text and data from PDF documents using the Poppler C++ library.	GITHUB

Leave us a message!

ropensci.org/packages/

5 matches chem Done

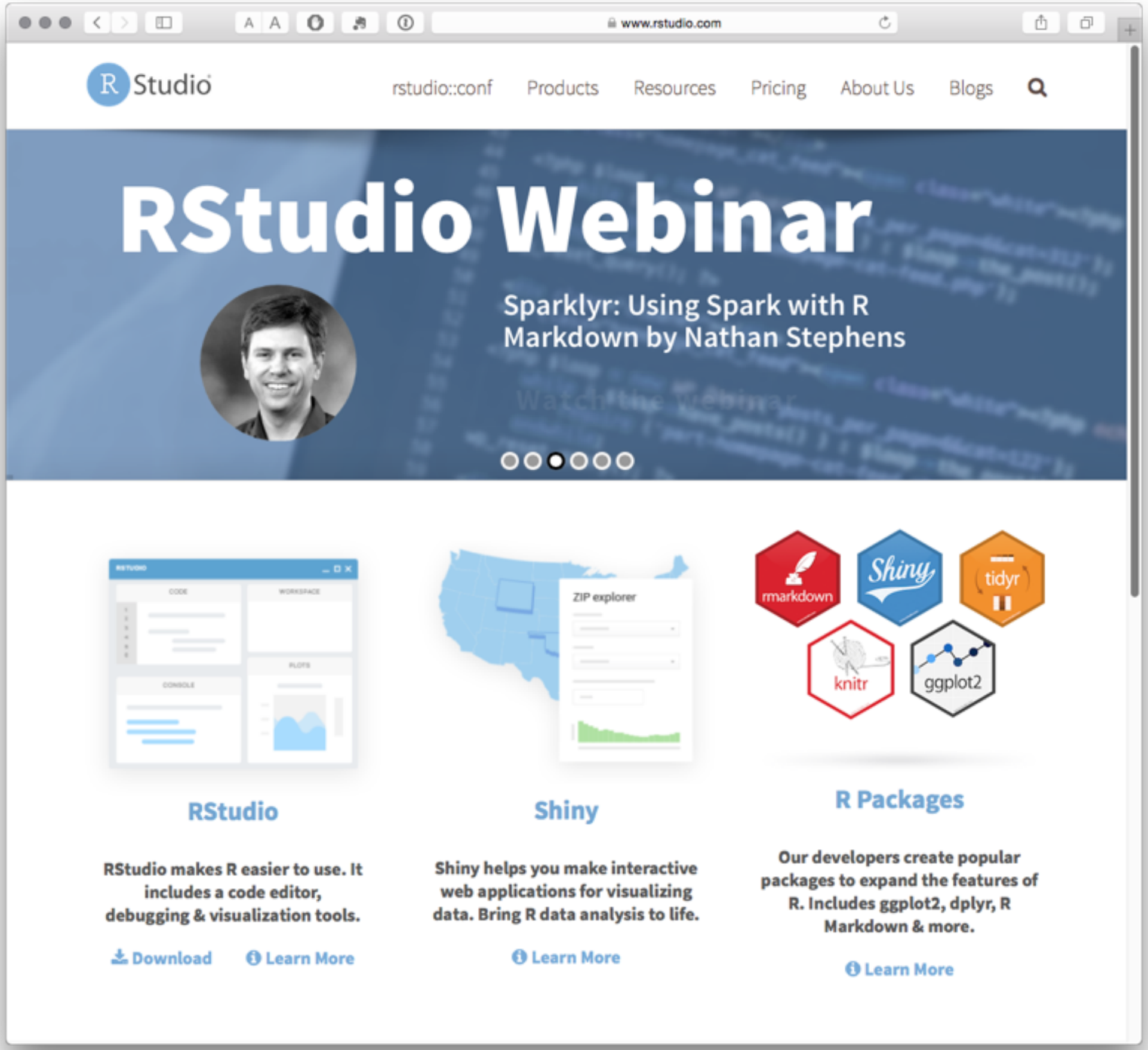
spocc	GBIF, iDigBio, iNaturalist, Ecoengine, AntWeb, eBird, and USGS's BISON.	GITHUB
stplanr	Functionality and data access tools for transport planning, including origin-destination analysis, route allocation and modelling travel patterns	CRAN GITHUB
taxize	Get taxonomic identifiers from many different databases; taxonomic synonyms; taxonomic hierarchies; retrieve common names, and more. Access to a dozen different sources	CRAN GITHUB
traits	A high level R package to acquire species trait data from many sources, including Polytraits, BETTYdb, NCBI, Phylomatic, and more.	CRAN GITHUB
treebase	Programmatic interface to Treebase - a database of phylogenetic trees	CRAN GITHUB
USAboundaries	Historical boundaries of the United States. Map the United States (or the colonies that became the United States) on any date from 1629 to 2000. Contains both county and state/territory level polygons.	CRAN GITHUB
webchem	webchem is a R package to retrieve chemical information from many sources. Currently includes: Chemical Identifier Resolver, ChemSpider, PubChem, and Chemical Translation Service.	CRAN GITHUB

Leave us a message!

Accessing APIs for the scientific literature using RStudio

- We will be using R and RStudio to access APIs.
- The reason for this is that RStudio is a very good platform for accessing multiple data sources and turning the data into data tables that we can export to use in other software.
- In the rest of the session we will:
  - Get started with RStudio
  - Access scientific literature from PLOS
  - Briefly explore other data sources such as rcrossref and fulltext.
- First we need to get up to speed with RStudio.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION



http://www.rstudio.com

www.rstudio.com/products/rstudio/

RStudio

rstudio::conf Products Resources Pricing About Us Blogs

Take control of your R code

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. Click here to see more RStudio features.

RStudio is available in open source and commercial editions and runs on the desktop (Windows, Mac, and Linux) or in a browser connected to RStudio Server or RStudio Server Pro (Debian/Ubuntu, RedHat/CentOS, and SUSE Linux).

CLICK HERE TO SEE ADDITIONAL FEATURES

Desktop

Run RStudio on your desktop

RStudio Desktop ›

Server

Centralize access and computation

RStudio Server ›

RStudio Desktop

Open Source Edition	Commercial License

- Access RStudio locally
- Syntax highlighting, code completion,

https://www.rstudio.com/products/rstudio/

www.rstudio.com/products/rstudio/#Desktop

RStudio

rstudio::conf | Products | Resources | Pricing | About Us | Blogs

RStudio Desktop

	Open Source Edition	Commercial License
Overview	• Access RStudio locally • Syntax highlighting, code completion, and smart indentation • Execute R code directly from the source editor • Quickly jump to function definitions • Easily manage multiple working directories using projects • Integrated R help and documentation • Interactive debugger to diagnose and fix errors quickly • Extensive package development tools	All of the features of open source; plus: • A commercial license for organizations not able to use AGPL software • Access to priority support
Support	Community forums only	• Priority Email Support • 8 hour response during business hours (ET)
License	AGPL v3	RStudio License Agreement
Pricing	Free	$995/year
	DOWNLOAD RSTUDIO DESKTOP	BUY NOW

RStudio Server

	Open Source Edition	Commercial License
	• Access via a web browser	All of the features of open source; plus:

https://www.rstudio.com/products/rstudio/#Desktop

www.rstudio.com/products/rstudio/download/

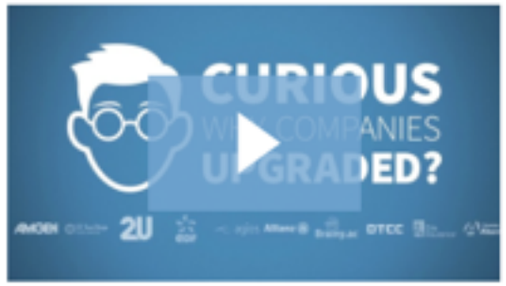
RStudio

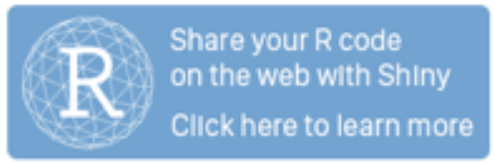
rstudio::conf Products Resources Pricing About Us Blogs

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

If you run R on a Linux server and want to enable users to remotely access RStudio using a web browser please download RStudio Server.

Do you need support or a commercial license? Check out our commercial offerings





RStudio Desktop 1.0.44 — Release Notes

RStudio requires R 2.11.1 (or higher). If you don't already have R, you can download it here.

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 1.0.44 - Windows Vista/7/8/10	81.9 MB	2016-11-01	7ccedc36c1f0a0861393763cfbe1c61d
RStudio 1.0.44 - Mac OS X 10.6+ (64-bit)	71.1 MB	2016-11-01	32256c7ac6d6597192a1bafa56a2747f
RStudio 1.0.44 - Ubuntu 12.04+/Debian 8+ (32-bit)	85.4 MB	2016-11-01	5f7fb95ee727606e9779af7bfe6fc6a8
RStudio 1.0.44 - Ubuntu 12.04+/Debian 8+ (64-bit)	92 MB	2016-11-01	074b7d3336ad07e32d10553f9669194a
RStudio 1.0.44 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	84.6 MB	2016-11-01	a5b203d482c6ab9ab77c5daf3fad5b8a
RStudio 1.0.44 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	85.6 MB	2016-11-01	bdc2cf31061d393a5d6626329f19bd6f

Zip/Tarballs

Zip/tar archives	Size	Date	MD5
RStudio 1.0.44 - Windows Vista/7/8/10	117.5 MB	2016-11-01	155dde44732ec850a76f16fc2561
RStudio 1.0.44 - Ubuntu 12.04+/Debian 8+ (32-bit)	86.1 MB	2016-11-01	faf837b58c68229748e591ea5e92
RStudio 1.0.44 - Ubuntu 12.04+/Debian 8+ (64-bit)	93.1 MB	2016-11-01	5bc32b5a9ce4ba657b40e687d4b8
RStudio 1.0.44 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	85.3 MB	2016-11-01	cad59eb6d9f09d391209b76549fb
RStudio 1.0.44 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	86.6 MB	2016-11-01	108e856153da246d5dc4558c4054

Source Code

A tarball containing source code for RStudio v1.0.44 can be downloaded from here

https://www.rstudio.com/products/rstudio/download/

1. Install R for your system from the link. 2. Download RStudio for your system.

~/Desktop/Brazil Workshop/tidy_data - RStudio

Go to file/function

tidy_data

Console ~/Desktop/Brazil Workshop/tidy_dat

```
R version 3.1.3 (2015-03-09) -- "Smooth Sid
ewalk"
Copyright (C) 2015 The R Foundation for Sta
tistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit
)

R is free software and comes with ABSOLUTEL
Y NO WARRANTY.
You are welcome to redistribute it under ce
rtain conditions.
Type 'license()' or 'licence()' for distrib
ution details.

  Natural language support but running in a
n English locale

R is a collaborative project with many cont
ributors.
Type 'contributors()' for more information
and
'citation()' on how to cite R or R packages
in publications.
```

Environment

Global Enviro

Files **Plots**

New Folder

Desktop

Name

..

.Rhisto

texput.

tidy_da

tidy_data.Rproj 204

New Project...

Open Project...

Open Project in New Window...

Close Project

tidy_data

OPS

gephi

google_api

open_refine

rgibif

lapply

shiny

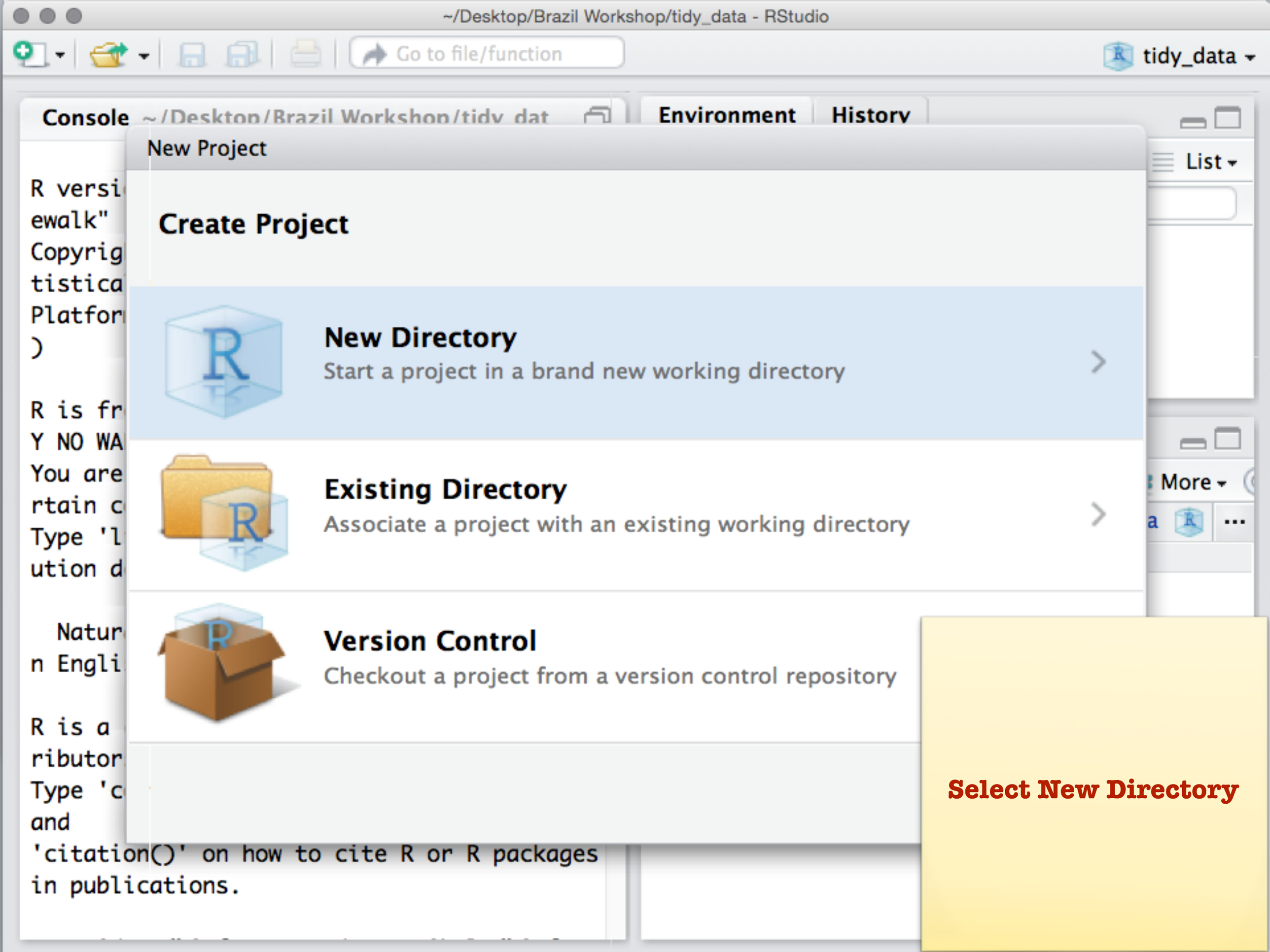
Regional_Workshop

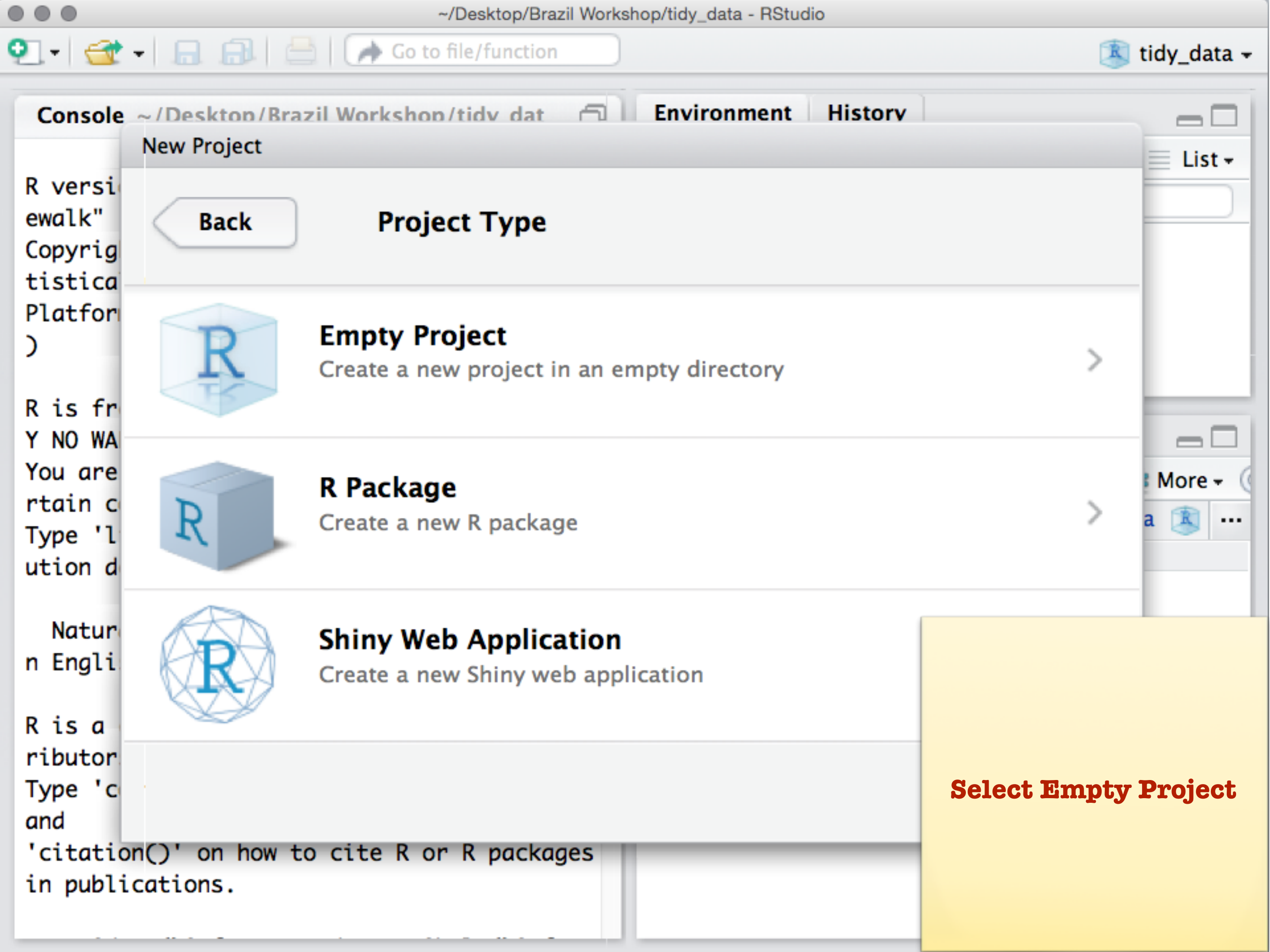
WORMS

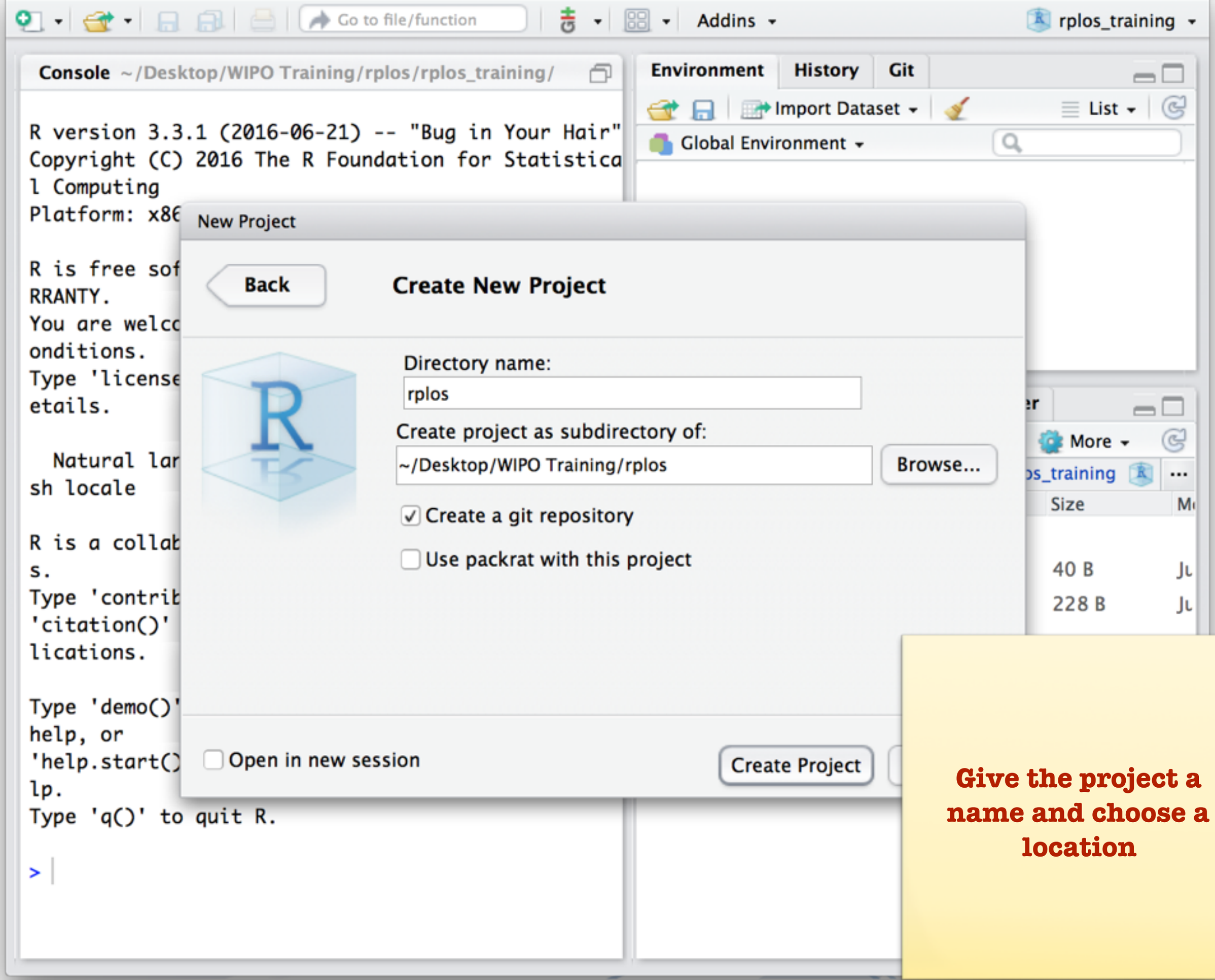
Clear P

Project

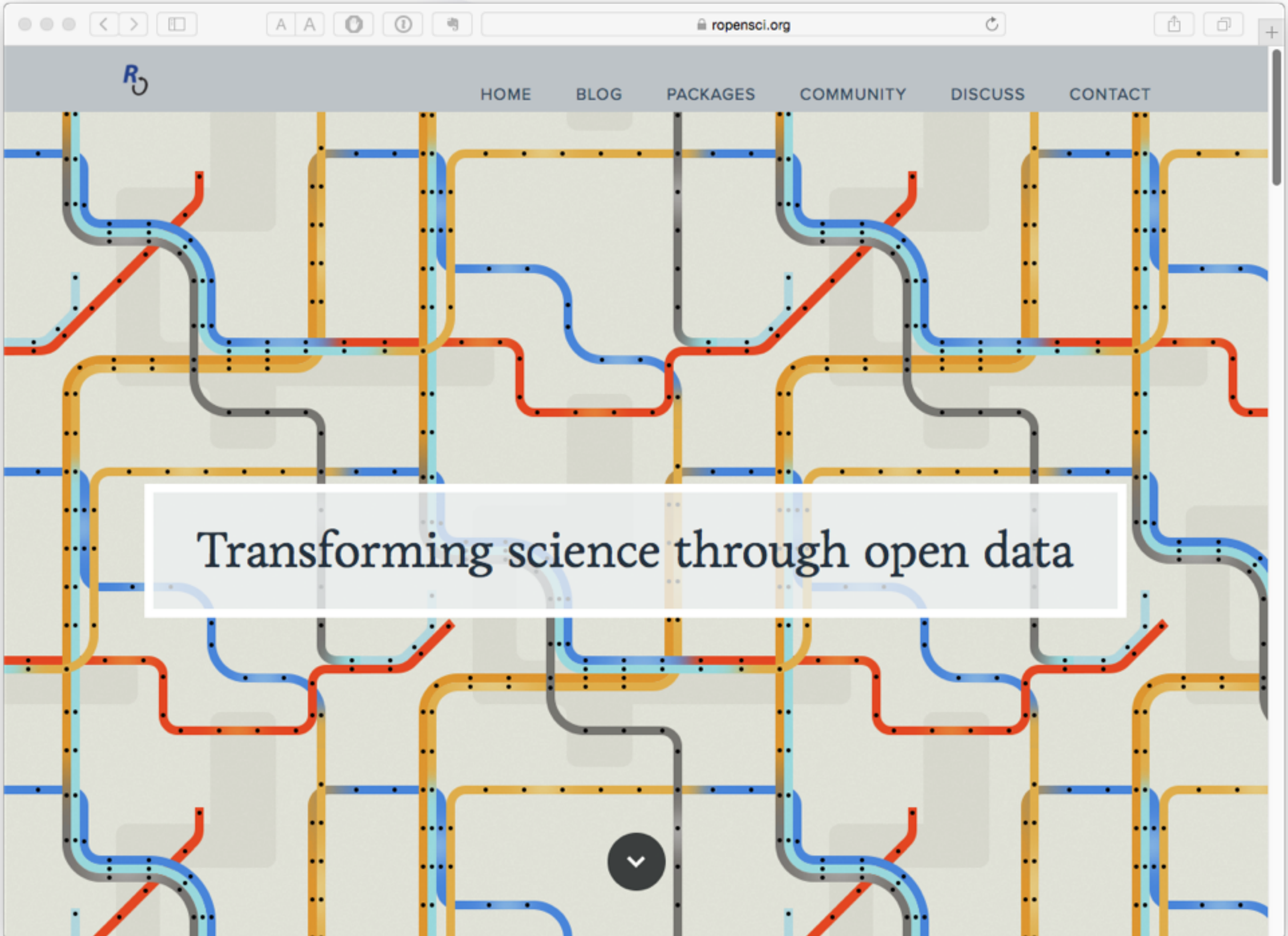
RStudio works with Projects. So let's create one


~/Desktop/Brazil Workshop/tidy_data - RStudio
Go to file/function
tidy_data
Console
Environment
History
New Project
Create Project
New Directory
Start a project in a brand new working directory
Existing Directory
Associate a project with an existing working directory
Version Control
Checkout a project from a version control repository
List
More
Select New Directory


~/Desktop/Brazil Workshop/tidy_data - RStudio
Go to file/function
tidy_data
Console
Environment
History
New Project
Back
Project Type
Empty Project
Create a new project in an empty directory
R Package
Create a new R package
Shiny Web Application
Create a new Shiny web application
List
More
Select Empty Project



Give the project a name and choose a location



http://ropensci.org

ropensci.org/packages/

HOME BLOG PACKAGES COMMUNITY DISCUSS CONTACT

rOpenSci packages

This is a complete list of all available rOpenSci packages. Packages are grouped by ones that acquire data, full-text of journal articles, altmetrics, data-publication, focus on scalable and reproducibile computing, data visualization, or geospatial work. Packages with a CRAN sign are stable versions that you can quickly install from your nearest mirror using `install.packages("PACKAGE_NAME")`. Others are in various stages of development (bleeding edge packages are not listed here) and you can learn more by following our GitHub organization page, and our GitHub organization for bleeding edge projects. All of our software packages are open source. Please see package description files for more information on specific licenses. We also have a package status dashboard that updates periodically.

Many of our packages are community-contributed. If you are interested in contributing a package, please visit our onboarding repository for details.

Data Publication | Data Access | Literature | Altmetrics | Scalable & Reproducible Computing | Databases | Data Vizualization | Data Tools | Geospatial

Leave us a message!

http://ropensci.org/packages/

ropensci.org/packages/

Literature

Packages that interface with journal full-text or metadata

Package	Description	Details
aRxiv	Access to the aRxiv preprint repository	CRAN GITHUB
bmc	An R interface to Biomed Central (BMC) search API and full text XML. IN EARLY DEVELOPMENT	CRAN GITHUB
csl	Explore CSL styles and locales. IN EARLY DEVELOPMENT	CRAN GITHUB
fulltext	An high level R interface to many different sources for full text (or nearly) scholarly texts, including PLOS, PMC, PeerJ, eLife, arXiv, and many more.	CRAN GITHUB
IEEER	Interface to the IEEE Xplore Gateway. IN EARLY DEVELOPMENT	CRAN GITHUB
oai	Work with OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) services, a protocol developed by the Open Archives Initiative.	CRAN GITHUB
pdftools	A fast and portable package to extract text and data from PDF documents using the Poppler C++ library.	GITHUB

Leave us a message!

ropensci.org/packages/

5 matches chem Done

spocc	GBIF, iDigBio, iNaturalist, Ecoengine, AntWeb, eBird, and USGS's BISON.	GITHUB
stplanr	Functionality and data access tools for transport planning, including origin-destination analysis, route allocation and modelling travel patterns	CRAN GITHUB
taxize	Get taxonomic identifiers from many different databases; taxonomic synonyms; taxonomic hierarchies; retrieve common names, and more. Access to a dozen different sources	CRAN GITHUB
traits	A high level R package to acquire species trait data from many sources, including Polytraits, BETTYdb, NCBI, Phylomatic, and more.	CRAN GITHUB
treebase	Programmatic interface to Treebase - a database of phylogenetic trees	CRAN GITHUB
USAboundaries	Historical boundaries of the United States. Map the United States (or the colonies that became the United States) on any date from 1629 to 2000. Contains both county and state/territory level polygons.	CRAN GITHUB
webchem	webchem is a R package to retrieve chemical information from many sources. Currently includes: Chemical Identifier Resolver, ChemSpider, PubChem, and Chemical Translation Service.	CRAN GITHUB

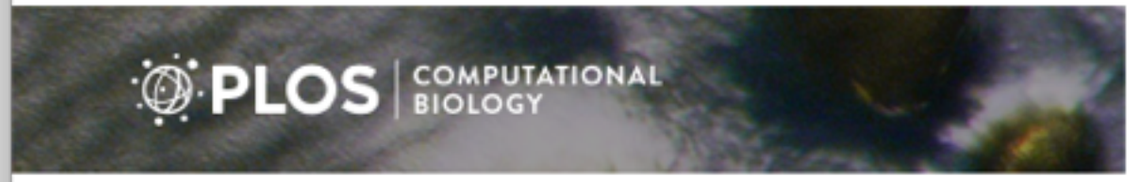
Leave us a message!

www.plos.org

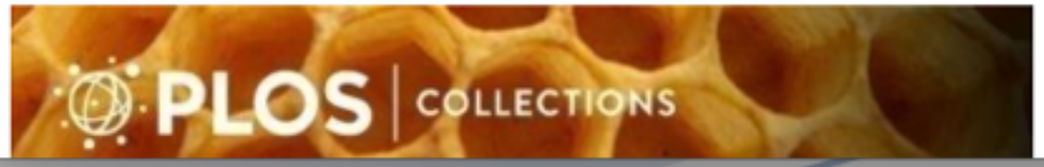
Explore PLOS Journals

160,000+ peer-reviewed articles are free to access, reuse and redistribute.

Anytime, Anywhere.



Dive In to More PLOS Content



https://www.plos.org/publications

PLOS ONE: Biological Diversit

journals.plos.org/plosone/article?id=10.1371/journal.pone.0078737

plos.org create account sign in

PLOS | ONE

Publish About Browse Search

advanced search

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Biological Diversity in the Patent System

Paul Oldham, Stephen Hall, Oscar Forero

Published: November 12, 2013 • http://dx.doi.org/10.1371/journal.pone.0078737

54 Save | 9 Citation | 9,787 View | 80 Share

Article Authors Metrics Comments Related Content

Download PDF

Print Share

CrossMark

Abstract
Introduction
Methods
Results
Discussion
Conclusion

Abstract

Biological diversity in the patent system is an enduring focus of controversy but empirical analysis of the presence of biodiversity in the patent system has been limited. To address this problem we text mined 11 million patent documents for 6 million Latin species names from the *Global Names Index* (GNI) established by the Global Biodiversity Information Facility (GBIF) and Encyclopedia of Life (EOL). We identified 76,274 full Latin species names from 23,882

Included in the Following Collection

Text Mining

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0078737

www.crossref.org

crossref.org

Contact | Members Area

ABOUT CROSSREF | FOR PUBLISHERS | FOR LIBRARIES | FOR AFFILIATES | FOR RESEARCHERS

Meetings & News

- Crossref Blog
- 2016 Annual Meeting
- Join Crossref
- Sign Up for a Webinar!
- Crossref Videos
- Online Payment Portal

Technical Resources

- Crossref Support
- Crossref Labs
- Report a DOI problem
- Web deposit form
- Simple Text Query
- XML Tools
- Browsable title list

CrossRef Services

- Cited-by Linking
- Similarity Check (formerly CrossCheck)
- CrossMark
- Crossref Metadata Services
- Crossref Text and Data Mining
- Funding Data
- ORCID Auto-Update
- Join Crossref
- Crossref Service Providers

Metadata Search

drones

Search Crossref's database of 80 million records for authors, titles, DOIs, ORCIDs, ISSNs, funders, license URIs, etc. You can even paste entire references into the search box and discover their DOIs.

Help! Example queries | API Documentation

We are Crossref, a not-for-profit membership organization for scholarly publishing working to make content easy to find, link, cite and assess. We do it in five ways: rallying the community; tagging the metadata; running a shared infrastructure; playing with new technology; and making tools and services to improve research communications.

It's as simple—and as complicated—as that.

80110857

registered CrossRef DOI links

millions of links

Google™ Custom Search

Search this Site

Crossref

Follow us on:

type crossref into google then drones

http://www.crossref.org

search.crossref.org/?q=drones

Crossref

drones

Status API Help Sign in

TYPE

- Journal Article (540)
- Chapter (115)
- Conference Paper (85)
- Dataset (10)
- Book (8)
- Other (3)
- Component (2)
- Monograph (2)
- Report (2)
- Section (2)

YEAR

- 2015 (198)
- 2016 (108)
- 2014 (82)
- 2013 (70)
- 2012 (21)
- 2011 (19)
- 2009 (16)
- 2010 (12)
- 2008 (11)
- 1978 (7)

PUBLICATION

- Apidologie (38)
- Journal of Apicultural Research (27)
- SSRN Electronic Journal (25)

SORT BY: **RELEVANCE** PUBLICATION YEAR

PAGE 1 OF 770 RESULTS

drones-scaling-new-heights

Section published in SAGE Business Researcher

http://dx.doi.org/10.1177/237455680202.n7 Actions

Drones, jihad and justice

Chapter published in Wars of Terror

http://dx.doi.org/10.5040/9781474217095.0008 Actions

Drones, Resistance and Countersurveillance

Chapter published 24 Dec 2015 in Drones and Unmanned Aerial Systems on pages 243 to 266

Authors: Aleš Završnik

http://dx.doi.org/10.1007/978-3-319-23760-2_11 Actions

Theorizing Drones and Droning Theory

Chapter published 24 Dec 2015 in Drones and Unmanned Aerial Systems on pages 21 to 43

Authors: Mark Andrejevic

http://dx.doi.org/10.1007/978-3-319-23760-2_2 Actions

Drones

Chapter published in Opposing Perspectives on the Drone Debate

Authors: Steven Levine

http://dx.doi.org/10.1057/9781137432636.0016 Actions

europepmc.org

Sign in or create an account

Europe PMC

About | Tools | Developers | Help | Europe PMC plus

Search worldwide, life-sciences literature

Search | Advanced Search

E.g. "breast cancer" HER2 Smith J

Search more than abstracts

- **Abstracts** (31.1 million, including 26 million from PubMed)
- **Full text articles** (3.7 million)
- **Patents** (4.2 million)
- **Agricola records** (618,700)
- **NHS clinical guidelines** (760)

Link to public databases

Explore protein, gene, species and disease records **directly from articles**:

- UniProt
- Protein Data Bank (PDBe)
- European Nucleotide Archive (ENA)
- Wikipedia and other lay summaries

Get credit for your work

ORCID is a unique identifier for researchers which distinguishes you from every other researcher, and makes it easier to find your work.

Use our claiming tool to link your Europe PMC articles to

Feedback

https://wipo-analytics.github.io/scientific-literature-with-rplos.html

12.10 Selecting applicants using ...

12.11 Generating IPC Tables

12.12 Creating an infographic in i...

Using APIs

13 Scientific Literature with Rplos

13.1 Introduction

13.2 Install R and RStudio

13.3 Create A Project

13.4 Install Packages

13.5 Key functions in rplos

13.6 Data Fields in rplos

13.7 Basic Searching using sear...

13.8 Limit by journal

13.9 Obtaining the number of rec...

13.10 Writing the results and usin...

13.11 Proximity Searching

13.12 Searching Using Multiple P...

13.13 Tidying and Organising the ...

13.14 Filling Blank Spaces

13.15 Converting Dates

13.16 Adding columns

13.17 Add a count

13.18 Remove a column

13.19 Arranging the Data

Chapter 13 Scientific Literature with Rplos

13.1 Introduction

In this chapter we look at the use of the `rplos` package from rOpenSci to access the scientific literature from the Public Library of Science using the PLOS Search API.

The Public Library of Science (PLOS) is the main champion of open access peer reviewed scientific publications and has published somewhere in the region of 140,000 articles. These articles are a fantastic resource. PLOS includes the following titles.

- PLOS ONE
- PLOS Biology
- PLOS Medicine
- PLOS Computational Biology
- PLOS Genetics
- PLOS Pathogens
- PLOS Neglected Tropical Diseases
- PLOS Clinical Trials ()
- PLOS Collections (collections of articles)

PLOS is important because it provides open access to the full text of peer reviewed research. For researchers interested in working with R, `rplos` and its bigger sister package, the rOpenSci `fulltext` package are very important tools for gaining access to research.

This article is part of work in progress for the WIPO Manual on Open Source Patent Analytics. The Manual is intended to introduce open source analytics tools to patent researchers in developing countries and to be of wider use to the science and technology research community. An important part of patent research is being able to access and analyse the scientific literature

https://wipo-analytics.github.io/

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Switch to RStudio

```
library(rplos)
library(readr)
library(plyr)  # load before dplyr to avoid errors
library(dplyr)
library(tidyr)
library(stringr)
library(tm)
library(XML)
```

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

searchplos()

- ?searchplos
- **p <- searchplos(q = "pizza", fl = c("id","publication_date", "title", "abstract"), limit = 20)**
- p
- class(p)
- str(p)
- p$meta$numFound

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Selecting data in the results

- **dat <- p$data**
- Write the data to a file.
- **write_csv(dat, “dat.csv”)**

Searching with multiple search terms

- The way this works is that rplos searches for each phrase separately (we will come back to that).

- Create a vector called **s** using single and double quotes ‘ and “” around the terms.

s <- c('"synthetic biology"', '"synthetic genomics"', '"synthetic genome"', '"synthetic genomes”')

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Find the number of records (see file plos_records)

```
plos_records <- function(q) {
  library(plyr) # for ldply
  library(dplyr) # for pipes, select and filter
    lapply(q, function(x) searchplos(x, limit = 0)) %>%
    ldply("[[", 1) %>% # get meta from the lists
    select(numFound) %>% #  select numFound column of meta
    filter(numFound == max(numFound)) %>% # filter on max
numFound
      print() # print max value of numFound
}
```

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

A lot easier to use than it looked.

```
res <- plos_records(s) # 1,119
```

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Now we know how many results we want but need some more code to get them and bring back a table.

```
results <- lapply(s, function(x) searchplos(x, fl = c('id',
'author', 'publication_date', 'title', 'abstract'), limit =
res)) %>% # note res in limit
    setNames(s) %>% # adds terms in s
    ldply("[[", 2)
```

Let it run and then:

```
View(results)
```

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Filter

	abstract	title
	So far, conservation scientists have paid little attentio...	Synthetic Biology and Conservation of Nature: Wicked...
..	We have created the Knowledgebase of Standard Biol...	Standard Biological Parts Knowledgebase
	Since it was launched in 2006, PLOS ONE has publish...	The PLOS ONE Synthetic Biology Collection: Six Years ...
	Synthetic biology presents a particularly revealing exa...	Synthetic Biology: Caught between Property Rights, t...
..		Correction: Transcriptome-Wide Mapping of Pea Seed ..
		Correction: Novel Functions and Regulation of Cryptic..
..	Vibrio fischeri is a typical quorum-sensing bacterium ...	Artificially Constructed Quorum-Sensing Circuits Are ...
..	Synthetic Biology Open Language (SBOL) Visual is a gr...	SBOL Visual: A Graphical Language for Genetic Designs
...	Yeast is an ideal organism for the development and a...	Rational Diversification of a Promoter Providing Fine-T..
...	Plant synthetic biology requires software tools to assi...	Rule-Based Design of Plant Expression Vectors Using ...
.	Background: Synthetic biological systems are currentl...	Eugene – A Domain Specific Language for Specifying ...
..	Background: Modularity is a crucial issue in the engin...	Bottom-Up Engineering of Biological Systems through...

owing 1 to 13 of 1,607 entries

Tidying up

- We have 1607 records and this was more than we expected. We can quickly find out where the duplicates are by counting the distinct dois in the id fields

```
dplyr::n_distinct(results, id)
```

- We can address this using spread

```
spread_results <- tidyr::spread(results, .id, id)
```

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Filter

	"synthetic biology"	"synthetic genome"
ked…	10.1371/journal.pbio.1001530	NA
	10.1371/journal.pone.0017005	NA
ars …	10.1371/journal.pone.0043231	NA
, t…	10.1371/journal.pbio.0050058	10.1371/journal.pbio
eed …	10.1371/annotation/f8467b75-4eef-4c44-ad20-eee3…	NA
ptic…	10.1371/journal.pone.0143463	NA
re …	10.1371/journal.pone.0104578	NA
igns	10.1371/journal.pbio.1002310	NA
e-T…	10.1371/journal.pone.0033279	NA
ng …	10.1371/journal.pone.0132502	NA
ng …	10.1371/journal.pone.0018882	NA
gh	10.1371/journal.pone.0039407	NA

2 of 1,232 entries

Write the results to file

- **readr::write_csv(spread_results, “spread_results.csv)**

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Going further with the fulltext package

- Search across a range of data source
- install.packages(“fulltext”)

library(fulltext)

```
pizza_search <- ft_search(query = "pizza", from =
c("plos", "crossref"))
```

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Access the results

- We need to subset into the data.frames in the results from the different data providers.

- **plos <- pizza_search$plos$data # 10 dois**
- **crossref <- pizza_search$crossref$data # 10 records, full data**
- **View(plos)**
- **View(crossref)**

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

ssion.Rmd* | pizza_search$plos$data | pizza_search$crossref$data

Filter

title	type	url
Pizza Brain	journal-article	http://dx.doi.org/10.1
Formation of Heterocyclic Amines in Salami and Ham …	journal-article	http://dx.doi.org/10.1
FTP PASV 'pizza thief' exploit	journal-article	http://dx.doi.org/10.1
YESTERDAY'S PIZZA	journal-article	http://dx.doi.org/10.1
Pizza perfect	journal-article	http://dx.doi.org/10.1
Pasta, Pizza – Pickel?	journal-article	http://dx.doi.org/10.1
Consumer Acceptance of Pizza and Pizza Crust Made	journal-article	http://dx.doi.org/10.1

Showing 1 to 7 of 10 entries

Console | R Markdown

~/Desktop/open_source_master/wipo_tutorials/

```
8  10.1371/journal.pone.0020822
9  10.1371/journal.pone.0058486
10 10.1371/journal.pone.0078660
> View(pizza_search$plos$data)
> View(pizza_search$crossref$data)
```

Linking patent data with literature packages

- The Lens extracts dois from patent data. Those dois can be used to search services such as crossref.
- In the repository there is a file called **pizza_1000.rda** that can be opened in R to reveal a Lens dataset.
- The dois can be extracted into **lenspizza_fulldoi.rda**
- Let's use the file called lens **lenspizza_doi10.rda**

- **library**(rcrossref) # load the library
- **ten_results <- cr_works(dois = lenspizza_doi10$doi) %>% .$data**

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

title	type	URL
Proteolytic and rheological changes during ageing of …	journal-article	http://dx.d
Der goldene Löffel	journal-article	http://dx.d
Untersuchungsmethoden für Brotgetreide, Mehl und …	journal-article	http://dx.d
D. J. De Renzo: Bakery Products — Yeast Leavened. 4…	journal-article	http://dx.d
The adjuvant OM-174 induces both the migration and…	journal-article	http://dx.d
Inflexal®V a trivalent virosome subunit influenza vacci…	journal-article	http://dx.d
New technology platforms in the development of vac…	journal-article	http://dx.d
Imiquimod and the imidazoquinolones: mechanism o…	journal-article	http://dx.d
Mutants of the Escherichia coli heat-labile enterotoxin…	journal-article	http://dx.d
Divergent synthetic nucleotide motif recognition patt…	journal-article	http://dx.d

Showing 1 to 10 of 10 entries

title	type	URL
A Coordination-Based Data Dissemination Protocol for…	proceedings-article	http:
A framework for targeting banner advertising on the I…	proceedings-article	http:
A multimodal consumer information server with IVR …	proceedings-article	http:
A multimodal consumer information server with IVR …	proceedings-article	http:
A multimodal consumer information server with IVR …	proceedings-article	http:
A multimodal consumer information server with IVR …	proceedings-article	http:
A multimodal consumer information server with IVR …	proceedings-article	http:
A multimodal consumer information server with IVR …	proceedings-article	http:
A multimodal consumer information server with IVR …	proceedings-article	http:
A multimodal consumer information server with IVR …	proceedings-article	http:
A multimodal consumer information server with IVR …	proceedings-article	http:
A multimodal consumer information server with IVR …	proceedings-article	http:
A novel design for a MMIC 180 degree phase shifter	proceedings-article	http:
A panorama based method of personal positioning on	proceedings-article	http:

Showing 1 to 13 of 3,384 entries

Conclusion

- APIs are a powerful way of gaining access to data in a form that we can use for patent analytics (tables, text files).
- Many data providers are creating APIs as a growing trend (and as an alternative to people scraping web pages)
- The trick is to provide easy to use and reusable functions that automate the generation of tables and processing.