

Web Data Collection with R API

Peter Meißner / 2016-02-29 – 2016-03-04 / ECPR WSMT

How APIs Work

Twitter API - API with authentication

How APIs Work

How APIs work

- ▶ **A**pplication **P**rogramming **I**nterface
- ▶ lots of web Services provide APIs to access their data and services (Twitter, Google, Facebook, Wikipedia, ...)
- ▶ ... aka **a preset and structured way of getting data** (or posting data, or do whatever the web service allows)
- ▶ frees us building our own scraper, provides legal access
- ▶ forces us to understand the way the API works
- ▶ but for several services ready made R packages exist, e.g.
 - ▶ <http://cran.r-project.org/web/views/WebTechnologies.html>
 - ▶ Web Analytics, Genes and Genomes, Sports, Social media, News, Images, Graphics, Videos, Music, Marketing, Maps, Literature, Metadata, Text, and Altmetrics, Governemnt, Google, Biology, Earth Science, Data Depots

examples

- ▶ Google Maps
 - ▶ <http://maps.googleapis.com/maps/api/directions/json?origin=Konstanz,Germany1&destination=Bamberg>
 - ▶ <https://developers.google.com/maps/documentation/directions/>
- ▶ GitHub
 - ▶ <https://api.github.com/users/petermeissner>
 - ▶ <https://developer.github.com/v3/>
- ▶ Twitter (fails, authentication needed)
 - ▶ https://api.twitter.com/1.1/statuses/user_timeline.json
 - ▶ <https://dev.twitter.com/overview/documentation>
- ▶ more APIs
 - ▶ <http://www.programmableweb.com/apis>
 - ▶ <http://cran.r-project.org/web/views/WebTechnologies.html>

Twitter API - API with authentication

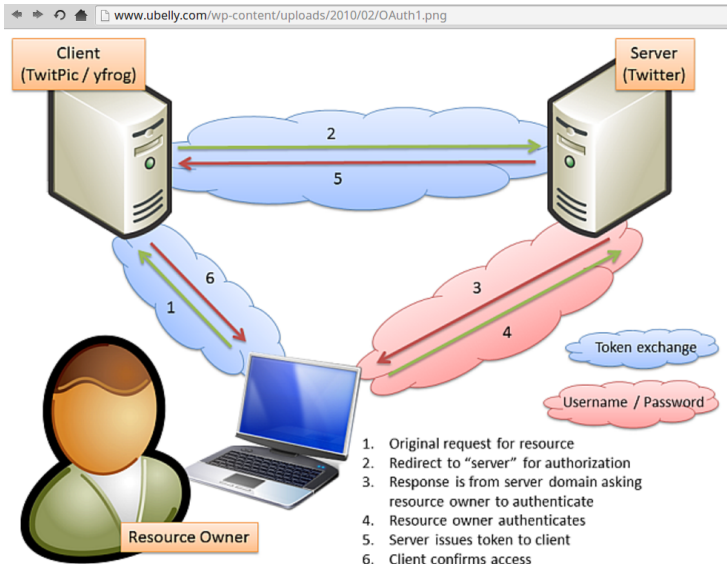
API with authentication

- ▶ provides API for tweeting, accessing tweets and user information
- ▶ more complex interface
- ▶ access needs OAuth authentication
 - ▶ get account / developer account
 - ▶ create/register application
 - ▶ use credentials to authorize

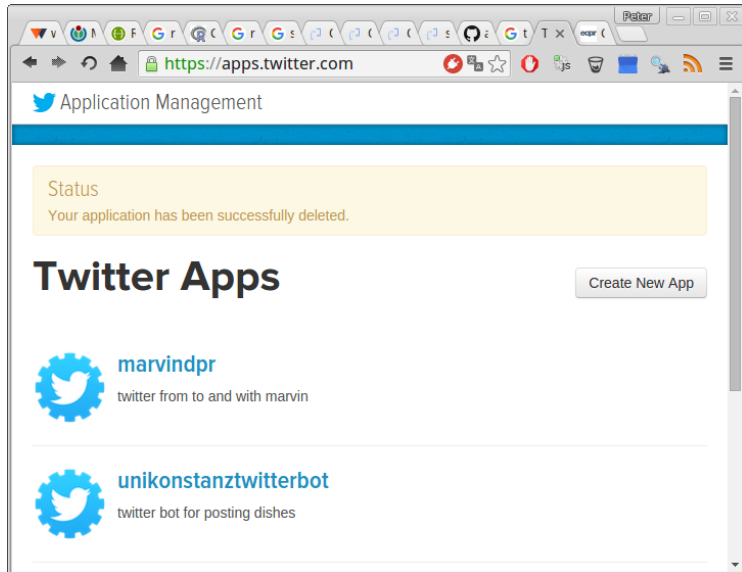
API with authentication

- ▶ *[httr]* has capabilities and some examples:
<https://github.com/hadley/httr/tree/master/demo>
- ▶ use an already written package
- ▶ ... twitterR package by Jeff Gentry (!must read!:
<http://geoffjentry.hexdump.org/twitterR.pdf>)

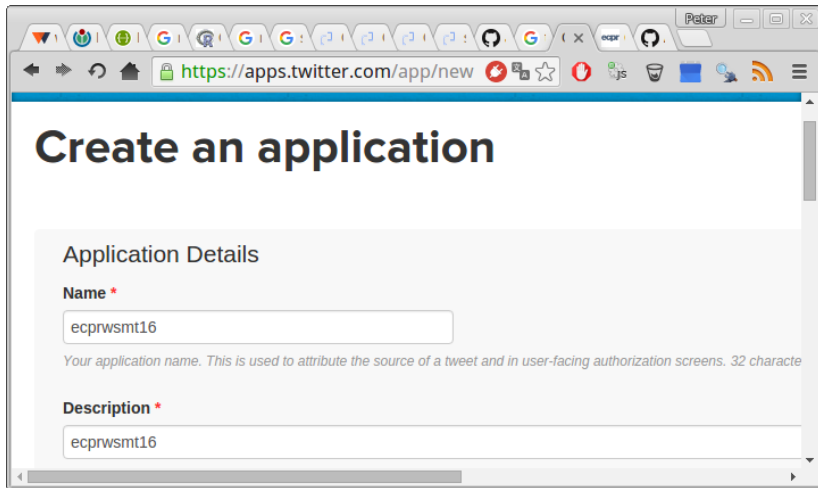
OAuth



twitter app



twitter app



A screenshot of a web browser window showing the Twitter 'Create an application' page. The browser's address bar displays the URL <https://apps.twitter.com/app/new>. The page title is 'Create an application'. Below the title, there is a section titled 'Application Details'. This section contains two required fields: 'Name *' and 'Description *'. Both fields have the text 'ecprwsmt16' entered. A small note below the 'Name' field states: 'Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters'.

Create an application

Application Details

Name *

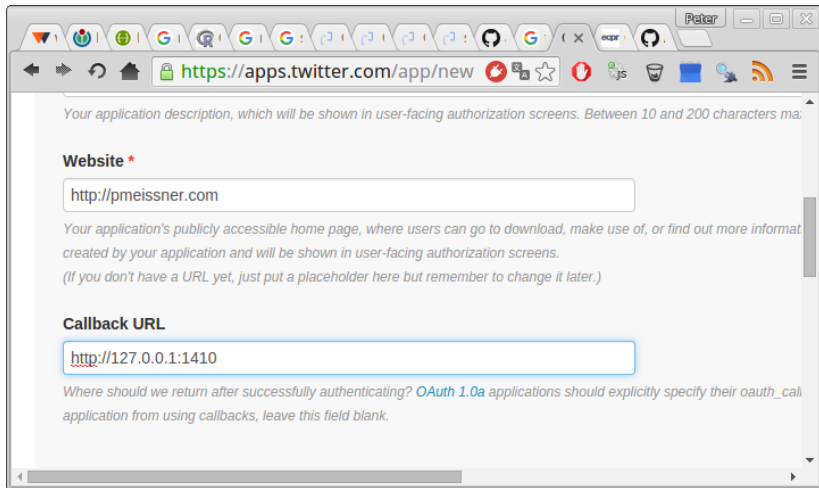
ecprwsmt16

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters

Description *

ecprwsmt16

twitter app



A screenshot of a web browser window showing the Twitter app creation page. The browser's address bar displays `https://apps.twitter.com/app/new`. The page content includes a description field, a 'Website' section with a text input containing `http://pmeissner.com`, and a 'Callback URL' section with a text input containing `http://127.0.0.1:1410`. The browser's tab bar shows multiple tabs, and the user's name 'Peter' is visible in the top right corner of the browser window.

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

`http://pmeissner.com`

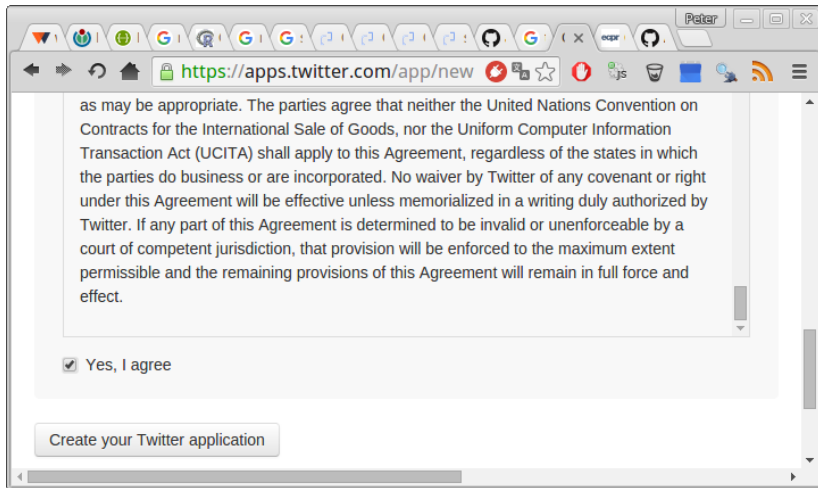
*Your application's publicly accessible home page, where users can go to download, make use of, or find out more information created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)*

Callback URL

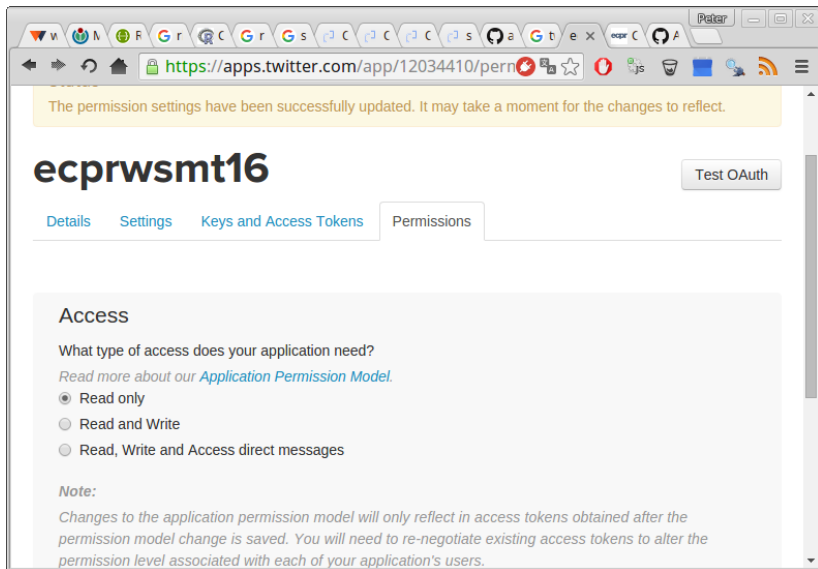
`http://127.0.0.1:1410`

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback application from using callbacks, leave this field blank.

twitter app



twitter app



The screenshot shows a web browser window with the URL <https://apps.twitter.com/app/12034410/permissions>. A yellow notification bar at the top states: "The permission settings have been successfully updated. It may take a moment for the changes to reflect." The app name "ecprwsmt16" is displayed prominently, with a "Test OAuth" button to its right. Below the app name are four tabs: "Details", "Settings", "Keys and Access Tokens", and "Permissions", with "Permissions" being the active tab. The main content area is titled "Access" and asks "What type of access does your application need?". It provides a link to "Read more about our Application Permission Model." and three radio button options: "Read only" (selected), "Read and Write", and "Read, Write and Access direct messages". A "Note:" section at the bottom explains that changes to the permission model only affect new access tokens and require re-negotiating existing ones.

The permission settings have been successfully updated. It may take a moment for the changes to reflect.

ecprwsmt16

Test OAuth

Details Settings Keys and Access Tokens Permissions

Access

What type of access does your application need?

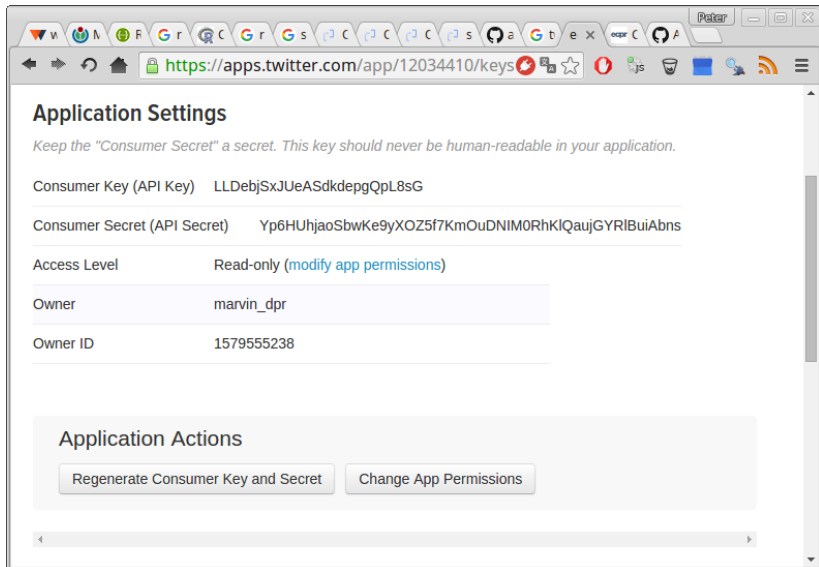
[Read more about our Application Permission Model.](#)

- ☒ Read only
- ☐ Read and Write
- ☐ Read, Write and Access direct messages

Note:

Changes to the application permission model will only reflect in access tokens obtained after the permission model change is saved. You will need to re-negotiate existing access tokens to alter the permission level associated with each of your application's users.

twitter app



The screenshot shows a web browser window with the address bar displaying `https://apps.twitter.com/app/12034410/keys`. The page title is "Application Settings". Below the title, a note states: "Keep the 'Consumer Secret' a secret. This key should never be human-readable in your application." The settings are organized into a table-like structure with labels on the left and values on the right. The "Consumer Key (API Key)" is `LLDebjSxJUeASdkdepgQpL8sG`. The "Consumer Secret (API Secret)" is `Yp6HUhjaoSbwKe9yXOZ5f7KmOuDNIM0RhKIQaujGYRIBuiAbns`. The "Access Level" is "Read-only (modify app permissions)". The "Owner" is `marvin_dpr`. The "Owner ID" is `1579555238`. Below the settings table, there is a section titled "Application Actions" containing two buttons: "Regenerate Consumer Key and Secret" and "Change App Permissions".

Application Settings

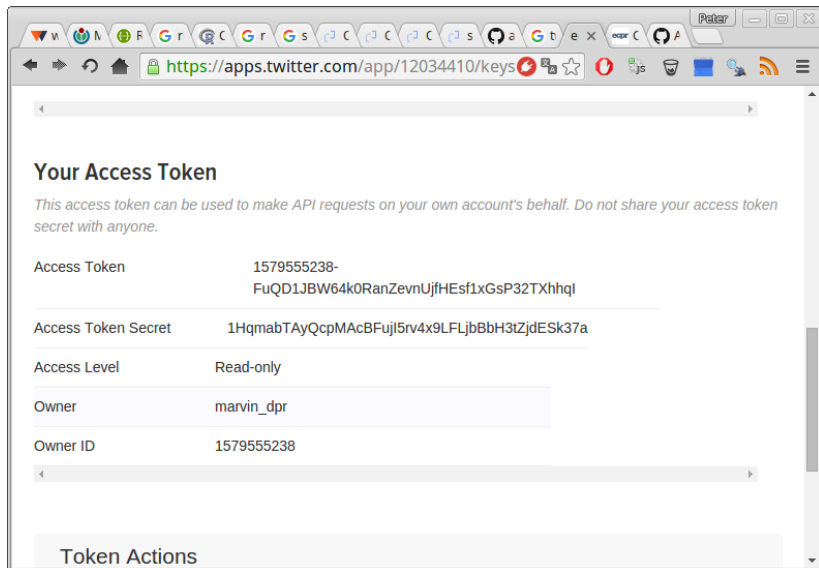
Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	LLDebjSxJUeASdkdepgQpL8sG
Consumer Secret (API Secret)	Yp6HUhjaoSbwKe9yXOZ5f7KmOuDNIM0RhKIQaujGYRIBuiAbns
Access Level	Read-only (modify app permissions)
Owner	marvin_dpr
Owner ID	1579555238

Application Actions

[Regenerate Consumer Key and Secret](#) [Change App Permissions](#)

twitter app



The screenshot shows a web browser window with the address bar displaying `https://apps.twitter.com/app/12034410/keys`. The page title is "Your Access Token". Below the title, a warning message states: "This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone." The page displays four key-value pairs for the access token:

Access Token	1579555238- FuQD1JBW64k0RanZevnUjfhESf1xGsP32TXhhql
Access Token Secret	1HqmabTAyQcpMAcBFujl5rv4x9LFLjbBbH3tjdESk37a
Access Level	Read-only
Owner	marvin_dpr
Owner ID	1579555238

At the bottom of the page, there is a section titled "Token Actions".

the twitter example

```
# packages
library(httr)
library(dplyr)
library(magrittr)
# credentials
cred_file <- "ecpr_wsmt_2016.credentials"
tmp      <- readLines(cred_file)
tmp
```

```
## [1] "twitter_api_key=LLDebjSxJUeASdkdepgQpL8sG"
## [2] "twitter_api_secret=Yp6HUhjaoSbwKe9yX0Z5f7KmOuDNIM0F"
## [3] "twitter_access_token=1579555238-FuQD1JBW64k0RanZevr"
## [4] "twitter_access_token_secret=1HqmabTAyQcpMAcBFujI5rv"
```

the twitter example

```
key = stringr::str_replace(  
  grep("twitter_api_key=", tmp, value = T),  
  "twitter_api_key=", "")  
  
secret = stringr::str_replace(  
  grep("twitter_api_secret=", tmp, value = T),  
  "twitter_api_secret=", "")  
  
token = stringr::str_replace(  
  grep("twitter_access_token=", tmp, value = T),  
  "twitter_access_token=", "")  
  
token_secret = stringr::str_replace(  
  grep("twitter_access_token_secret=", tmp, value = T),  
  "twitter_access_token_secret=", "")
```

the twitter example

```
twitter_token <-  
  Token1.0$new(  
    endpoint      = NULL,  
    params        = list(as_header = TRUE),  
    app           = oauth_app( "twitter", key, secret ),  
    credentials   = list(  
      oauth_token      = token,  
      oauth_token_secret = token_secret  
    )  
  )  
)
```

the twitter example

```
req <-  
  GET(  
    paste0(  
      "https://api.twitter.com/1.1/search/tweets.json",  
      "?q=%23wsmt16&result_type=recent&count=100"  
    ),  
    config(token = twitter_token)  
  )
```

the twitter example

```
tweets <-  
  req %>%  
  content("parsed") %>%  
  extract2("statuses") %>%  
  lapply(`[`, "text") %>%  
  unlist(use.names=FALSE)
```

the twitter example

```
tweets %>% grep("^RT ",. ,invert=TRUE, value=TRUE)
```

```
## [1] "Let it snow, let it snow, let it snow @ECPR #wsmt16"
## [2] "I know it's winter school but does it really have"
## [3] "If you're at the #wsmt16 don't miss the Brown Bag"
## [4] "The 2016 ECPR Winter School is now in full swing!"
## [5] "Lots of levels, lots of interesting projects in @ECPR"
## [6] "Welcome to over 400 participants and instructors at"
## [7] "Looking forward to participate in #wsmt16 course"
## [8] "#wsmt16 here I come. Two hours late because of a t"
## [9] "Just arrived for @ECPR Winter school in Bamberg! @"
## [10] "Troubleshooting LaTeX at the @ECPR #WSMT16 food ar"
## [11] "Temporary office for the week. Hallohhh Bamberg @ECPR"
## [12] "I am excited to be back at @BAGSS5 for @ECPR #wsmt16"
## [13] "On my way. Going slowly but steady with Diesel-pow"
## [14] "On my way to Bamberg for the #wsmt16 @ECPR"
## [15] "Preparing the Advanced Process-Tracing Methods Cou"
## [16] "Looking forward to today's course on #NVivo softwa
```