|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Revisió | Redactat per | Revisat per | Aprovat per | Data aprovació | Data publicació |
| 1 | Sergi Guevara | Núria Majó Crespo |  |  | Abril 2021 |

|  |
| --- |
| Registre de canvis del document |

|  |  |  |  |
| --- | --- | --- | --- |
| Revisió | Apartat | Data Modificació | Motiu del canvi |
|  |  |  |  |
|  |  |  |  |

|  |
| --- |
| Responsable del document: Oficina Tècnica de Dades Obertes |

Í N D E X

1. Introducció 2

1.1 Dades obertes 2

1.2 Objecte del document 2

2. RSocrata 3

2.1 Instal·lar el paquet 3

2.2 Funcions del paquet 3

2.2.1 Carregar les dades 3

2.2.2 Llistar els datasets disponibles 4

2.2.3 Sobreescriure un dataset 5

2.2.4 Convertir una variable calendar\_date de Socrata a una POSIX 5

2.2.5 Convertir el nom d’una variable a un nom que pugui utilitzar R 5

3. Definicions, acrònims i abreviatures 6

# Introducció

## Dades obertes

La solució de **Dades obertes de la Generalitat de Catalunya**, fonamentada en la plataforma *Socrata* ofereix diverses maneres de treballar amb les dades: descarregant les dades manualment del portal o bé mitjançant la connexió web amb un enllaç.

Una de les eines d’explotació de les dades és R (o la seva interfície gràfica del programa, R Studio), una eina molt bona per analitzar-les i que, a més, disposa d’un paquet específic amb funcions per accedir a dades de Socrata.

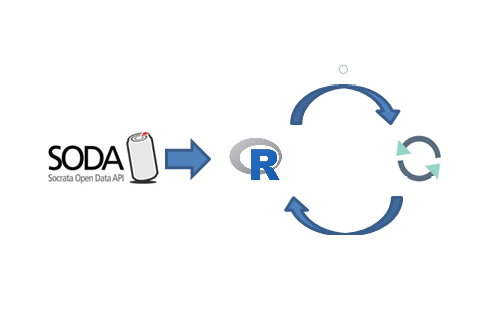
## Objecte del document

L’objectiu d’aquest document és mostrar els passos a seguir a R Studio per tal de poder obtenir dades de Socrata amb un enllaç, sense haver de descarregar el fitxer CSV (per Europa), i opcionalment després tornar a carregar-les a Socrata.

El mètode que mostrem en aquesta guia utilitza la connexió via API de Socrata. Això vol dir que no descarregarem les dades en cap moment, sinó que R estarà connectat “online” amb Socrata, obtenint la última versió de les dades cada cop que llencem la comanda de càrrega.

Aquest sistema ens permet treballar amb una eina professional a l’hora de realitzar tractament de dades.

Cal tenir en compte que generalment aquesta relació és unidireccional, ja que R agafa les dades de Socrata, però no podem sobreescriure les dades de Socrata a partir del que treballem a R, a menys que es tinguin els permisos necessaris per a poder modificar les dades de Socrata, amb la comanda específica per pujar els canvis realitzats.



# RSocrata

## Instal·lar el paquet

Primer de tot hem d’instal·lar el paquet específic amb les funcions per a llegir de Socrata a l’entorn de R on treballarem. S’ha d’instal·lar i, posteriorment, carregar el paquet al programari.



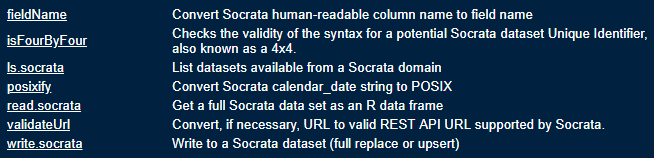
Aquesta seria la comanda per instal·lar el paquet



I a continuació carreguem el paquet al programa

## Funcions del paquet

Aquestes són totes les funcions que té el paquet Rsocrata amb una petita descripció del que fa cada funció.



### Carregar les dades

El primer pas és accedir al conjunt de dades amb la comanda **“read.socrata(“”)”**

Un cop tenim el paquet carregat i la comanda necessària per agafar el dataset, hem d’obtenir la URL especifica d’aquest conjunt de dades des de la pàgina del portal de dades obertes.

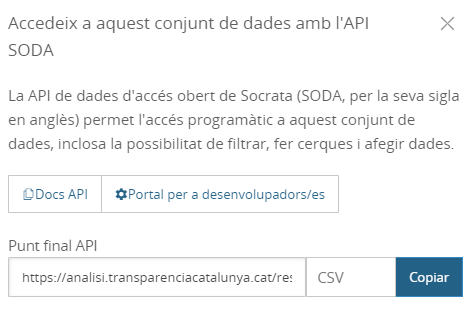
Aquesta URL és el punt d’accés API (API Access point) i te el format:

[http://analisi.transparenciacatalunya.cat/resource/<SOCRATA ID>.csv](http://analisi.transparenciacatalunya.cat/resource/%3cSOCRATA%20ID%3e.csv)

On “SOCRATA\_ID”,és l’ID del dataset al que volem connectar-nos, que sempre té un format de 4 dígits un guió i 4 dígits. (xxxx-xxxx).

El podem obtenir fàcilment des de la pàgina de metadades del portal amb l’opció “API” i seleccionant l’opció “CSV” i copiant-la al portapapers.





Un cop tinguem la URL, l’hem d’introduir a la comanda per connectar-se.



A part de carregar les dades de Socrata, és important assignar aquestes dades a una variable. En aquest cas, per exemple, s’assignaria a la variable “dades”.

A partir d’aquí, si es vol treballar amb les dades de Socrata, haurem d’utilitzar la variable a la qual li hem assignat el read.socrata.

Hi ha alguns datasets que són privats i es necessitarà proporcionar unes credencials d’accés. Per aquests datasets s’haurà d’introduir un correu i una contrasenya en la crida. La sintaxis serà:

**“read.socrata(“http://analisi.transparenciacatalunya.cat/resource/<xxxx-xxxx>.csv”, email, password)”**

### Llistar els datasets disponibles

Una altra opció és la de veure quins datasets hi ha disponibles, amb la comanda **"ls.socrata"**. Aquesta comanda mostra tots els datasets que conté Socrata a l'URL que s'hagi proporcionat. Cal posar entre cometes l'adreça del portal.

Es poden incorporar a la crida els diferents paràmetres que admet la cerca al catàleg de Socrata. Per exemple, per cercar els que tenen la paraula COVID:



Si es volgués buscar datasets amb una altra paraula clau, només caldria canviar la paraula “covid” per la que convingui.

Per més informació sobre l’API: [https://socratadiscovery.docs.apiary.io/#](https://socratadiscovery.docs.apiary.io/).

### Sobreescriure un dataset

Una altra funció del paquet RSocrata és la de sobreescriure un dataset. Aquesta funció permet tornar a Socrata el dataset que has estat utilitzant editat amb R. No obstant això, cal saber que Socrata està protegit amb contrasenya, així que no es podrà escriure si no s’especifica un nom i una contrasenya d’un usuari que tingui drets per a fer-ho.

La comanda és:

**“write.socrata(dataframe, dataset\_json\_endpoint, update\_mode, Email, password)”**

On cada paràmetre significa el següent:

* ***Dataframe:*** les dades amb les que s’ha treballat i que es vol pujar a Socrata.
* ***Dataset\_json\_endpoint:*** el punt d’accés API del dataset original on es vol escriure. És important haver agafat l’API en format JSON (no en CSV) ja que el format d’Excel encara no està habilitat en aquesta versió del paquet.
* ***Update\_mode***: Hi ha dues opcions: “**UPSERT**” si es vol actualitzar el dataset afegint-hi només els nous canvis o **“REPLACE”** per substituir el que hi ha a Socrata i carregar les noves dades.
* ***Email***: correu identificador de l’usuari que té permís per editar aquest dataset a Socrata.
* ***Password:*** La contrasenya associada a l’usuari.

Així doncs, un exemple seria:



### Convertir una variable calendar\_date de Socrata a una POSIX

La diferència bàsica entre una variable de tipus calendar\_date i una de POSIX és que el calendar\_date no treballa amb hores mentre que el POSIX, sí. És per això que segons quin tipus de dataset estigueu utilitzant, pot ser necessari tenir les hores. Per això serveix la comanda **"posixify()"**, permet passar una variable que només tingui la data a format POSIX que contingui la data i l'hora.

### Convertir el nom d’una variable a un nom que pugui utilitzar R

Un problema que podem tenir a l'hora de donar noms a variables és que tinguin un nom amb el qual R no pot treballar. Per això hi ha la comanda **"fieldName()"** que transforma el nom d’una variable que és llegible per un humà per un altre nom que el programa R pugui utilitzar sense problemes. D'aquesta manera, s’eliminaran els espais i accents del nom de les variables.

# Definicions, acrònims i abreviatures

* **Socrata**: És una companyia de programari en el núvol que proporciona serveis de dades obertes. Ajuden a les organitzacions del sector públic a millorar la transparència, l'atenció al ciutadà i la presa de decisions basada en fets mitjançant el lliurament de dades de manera eficient als ciutadans, els empleats i els desenvolupadors en una experiència fàcil d'usar en les interfícies web, mòbil i de màquina a màquina.
* **DS (Dataset) - Conjunt de dades** És una col·lecció de dades guardades en format taula en Socrata.
* **CSV**: Un arxiu de valors separats per comes *“Comma Separated Values”* (CSV) que emmagatzema les dades tabulars (números i text) en text pla. Cada línia del fitxer és un registre de dades. Cada registre es compon d'un o més camps, separats per comes. L'ús de la coma com a separador de camps és la font del nom per a aquest format d'arxiu.
* **R:** és un ambient de programació format per un conjunt d'eines molt flexibles que poden ampliar-se fàcilment mitjançant paquets, llibreries o definint les nostres pròpies funcions. Per això mateix, R es una eina estadística estable, confiable i de Vanguardia, ja que té actualitzacions permanentment.
* **SHP:** Arxiu de dades en format *ESRI Shapefile*. Originàriament és un format propietari de dades espacials desenvolupat per la companyia [ESRI](https://ca.wikipedia.org/w/index.php?title=ESRI&action=edit&redlink=1), però actualment s'ha convertit en format estàndard *de facto* per a l'intercanvi d'informació geogràfica. Es composa de diversos fitxers que se solen comprimir en un únic arxiu ZIP.
* **DEXI:** Departament d’Acció Exterior, Relacions Institucionals i Transparència
* **OGD:** Oficina de Governança de Dades
* **OTD:** Oficina Tècnica de Dades