



Laboratorio Nacional de Políticas Públicas



CENTRO DE INVESTIGACIÓN  
Y DOCENCIA ECONÓMICAS A.C.

# Datos Abiertos

(y descarga automatizada)

## Periodismo de Datos

### Marzo, 2021

**M.C. JORGE JUVENAL CAMPOS FERREIRA.**

Investigador Asociado.

Laboratorio Nacional de Políticas Públicas

CIDE

## Hoja de Ruta.

- 1. Revisión y discusión sobre datos abiertos.**
- 2. Revisión del concepto de datos estructurados y datos no-estructurados.**
- 3. Revisión de las funciones para hacer descargas (y descargas masivas) en R.**
- 4. Revisión del concepto de bucle.**
- 5. Ejercicio práctico, aplicando lo visto en (3) y (4) con datos del tipo (1)**

# Conceptos Básicos



“Filosofía y práctica que persigue que **determinados tipos de datos estén disponibles de forma libre para todo el mundo**, sin restricciones de derechos de autor, de patentes o de otros mecanismos de control”.

*Open Knowledge foundation*

# Datos Abiertos



[Los datos] son abiertos si satisfacen las condiciones siguientes:

1. **Disponibles y de fácil acceso.** (Disponible como un todo desde el principio, a un costo razonable de reproducción, preferentemente desde internet).
2. **Reutilizables y redistribuibles** (los datos deben ser provistos bajo términos que permitan reutilizarlos, redistribuirlos e integrarlos con otros conjuntos de datos).
3. **Facilitan la participación universal** (todos deben poder utilizar, reutilizar y redistribuir la información, sin discriminación en términos de esfuerzo, personas o grupos).

# Datos Abiertos (léanlo después)



## -Datos Abiertos

[Los datos] son abiertos si su forma de distribución satisface las condiciones siguientes:

- 1.. [Los datos] deben estar disponibles integralmente y a un coste de reproducción razonable, preferiblemente descargable de internet de manera gratuita. Los datos igualmente deben estar disponibles en una forma conveniente para ser modificables.
- 2.. La licencia no debe restringir a nadie la posibilidad de vender o distribuir los datos en sí mismos o formando parte de un paquete hecho de otros conjuntos de datos. Igualmente, la licencia no debe exigir un pago u otro tipo de cuota para esta venta o distribución.
- 3.. La licencia debe permitir hacer modificaciones y obras derivadas y debe permitir que estas sean distribuidas en las mismas condiciones que la obra original. La licencia puede imponer algún tipo de requerimiento referente al reconocimiento y a la integridad.
4. Se deben proporcionar los datos de tal manera que no haya ningún obstáculo tecnológico para ejecutar los actos mencionados anteriormente (formato de datos abiertos). Un formato de Datos abiertos es un formato que pueda ser utilizado sin imponer ninguna restricción (ni de uso monetario, ni de sistema operativo o de otro tipo).
- 5.. Los datos abiertos pueden exigir como condición de uso para la redistribución y la reutilización el reconocimiento a los contribuyentes y a los creadores de los datos. Este reconocimiento no debe imponerse de manera onerosa.
- 6.. La licencia puede requerir como condición que la base de datos, si es modificada o procesada para ser distribuida, tenga un nombre que indique que es otra base de datos diferente.
- 7.La licencia no debe discriminar a ninguna persona o grupo de personas.
8. La licencia no debe restringir a nadie hacer uso de la obra en un ámbito de trabajo específico. Por ejemplo, no puede restringir el uso de la obra en un negocio, o que esta sea utilizada para investigación militar.
9. Los derechos adjuntos a la obra deben aplicarse también a cualquier persona a quién le sea redistribuida sin necesidad de que esta ejecute una licencia adicional



## Beneficios

- Mejorar el **entendimiento acerca del buen (o mal) funcionamiento** de las políticas públicas que implementa un gobierno.
- **Incrementan la eficiencia gubernamental**
- **Dan herramientas al público** para involucrarse en la vigilancia de las políticas públicas.
- Para desarrolladores, y con las herramientas digitales adecuadas, **se pueden construir aplicaciones de interés público.**

Lectura recomendada:

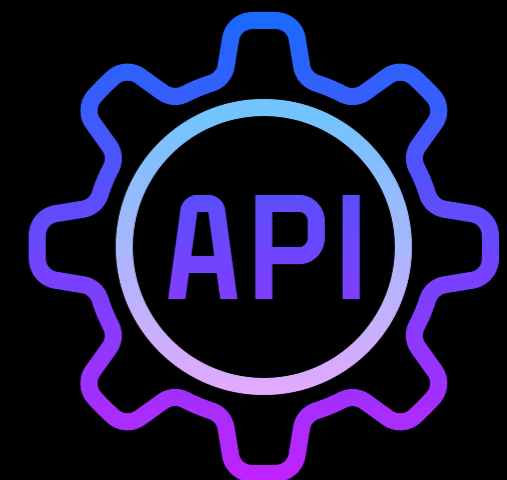
<https://opensource.com/government/16/5/why-open-data-matters>

# Acceso a los Datos Abiertos



Para acceder a los datos abiertos, tenemos 3 opciones:

1. A través de la **descarga de los datos en un sitio público** o la distribución de datos por parte de un tercero.
2. A través de la **lectura de los datos mediante un software de análisis de datos**.  
(Ejemplo, lo que hace `readr::read_csv()` con los documentos de internet).
3. A través de un **API** (Application Program Interface).



# **Discutamos antes los datos estructurados y no estructurados**

# Datos Estructurados y no estructurados

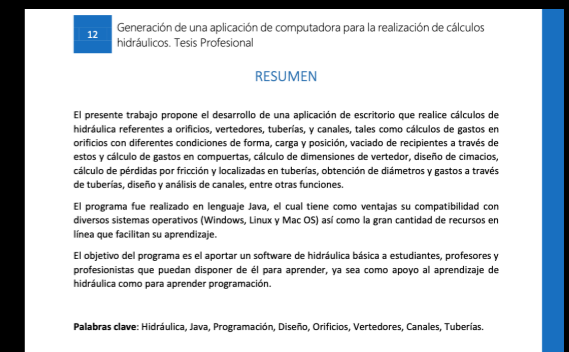
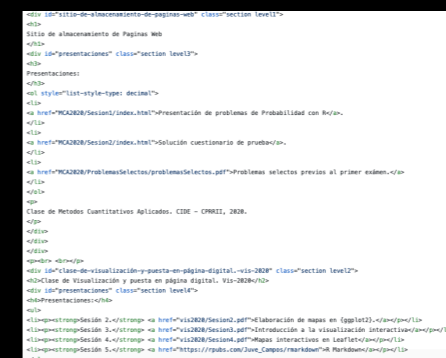
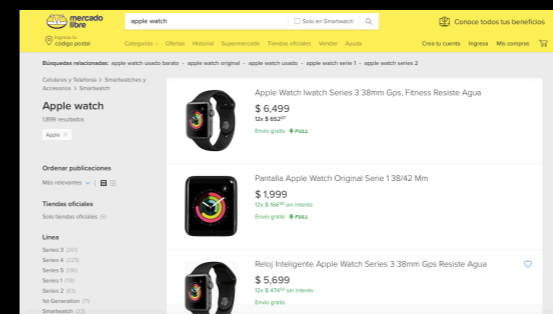


**Los datos estructurados son datos presentados en una forma predefinida. Típicamente (pero no exclusivamente) se presenta de forma tabular, en el que las columnas son variables y las filas observaciones.**

Entre estos datos tenemos los archivos de texto plano (\*.csv, \*.txt), los archivos de Office (\*.docx, \*.xlsx), los archivos que almacenan información geográfica (\*.shp, \*.geojson), entre otros, que tienen una estructura definida y son leídos por software de manera nativa.

**Los datos no-estructurados.** Son datos que no tienen un modelo predefinido de datos o bien no está presentado de forma ordenada. Típicamente son conjuntos de texto, imágenes, números, fechas que no están organizados de forma sistemática.

Entre estos se encuentran los \*.pdfs (algunos), los \*.pdfs fotografías, así como la información disponible en páginas de internet (catálogos, textos, etc).



Best departing flights				
Total price includes taxes + fees for 1 adult. Additional bag fees and other fees may apply.				
14:10 - 04:45 <sup>-1</sup>	24 h 35 m	2 stops	MX\$19,182	round trip
American, JAL	MEX - HND	DFW, LAX		
18:05 - 15:00 <sup>-2</sup>	30 h 55 m	1 stop	MX\$19,193	round trip
American, JAL	MEX - NRT	14 h 25 m DFW		
13:45 - 14:10 <sup>-3</sup>	34 h 25 m	1 stop	MX\$23,231	round trip
Delta	MEX - HND	17 h 20 m DTW		
Prices are currently typical for your trip.				
Other departing flights				
18:05 - 15:35 <sup>-1</sup>	55 h 30 m	2 stops	MX\$18,706	round trip
American, JAL	MEX - HND	DFW, ORD		
18:05 - 15:35 <sup>-1</sup>	55 h 30 m	2 stops	MX\$18,706	round trip
American, JAL	MEX - HND	DFW, ORD		
18:05 - 15:35 <sup>-1</sup>	55 h 30 m	2 stops	MX\$18,706	round trip
American, JAL	MEX - HND	DFW, ORD		

## Datos estructurados



**Los datos estructurados.**  
Empaquetados y listos para utilizarse.

## Datos no estructurados



**Los datos no-estructurados.**  
La información esta ahí, pero hay que extraerla y empaquetarla.

# Descarga de datos a través de R y RStudio.

Descarga automatizada 

# Descarga de datos a través de R y RStudio.

## Ventajas:

- En ciertos casos, te permite hacer las descargas más rápido que si hicieras el trabajo de manera manual
- Es más desafiante intelectualmente 🧐
- Permite a los lectores reproducir tu trabajo
- Permite solo compartir el código, sin tener que subir todos los archivos, dejando la descarga al lector

# Funciones a repasar

- `str_c()`, para concatenar texto.
- `curl::curl_download()`, para descargar archivos de internet (principalmente \*.zip y \*.xlsx)
- `zip::unzip()`, para deszippear o descomprimir archivos de internet.
- Función `for` para realizar bucles o ciclos.

# curl::curl\_download()

Función que sirve para bajar archivos de internet.

## curl::curl\_download()

Download file to disk

### Description

Libcurl implementation of `C_download` (the "internal" download method) with added support for https, ftps, gzip, etc. Default behavior is identical to `download.file`, but request can be fully configured by passing a custom `handle`.

### Usage

```
curl_download(url, destfile, quiet = TRUE, mode = "wb", handle =  
new_handle())
```

### Argumentos importantes

- \* **url**, la url donde se encuentra el archivo a descargar.
- \* **destfile**, como se va a llamar el archivo, una vez que lo descargue en la compu.

# **curl::curl\_download()**

Función que sirve para bajar archivos de internet.

## **RCurl::curl\_download()**

### **Ejemplo de uso:**

```
curl::curl_download(url = "https://www.inegi.org.mx/contenidos/programas/intercensal/2015/tabulados/14\_vivienda\_mor.xls",  
                    destfile =  
                      "01_Datos/Datos Censo/HogaresMorelos.xls")
```

Descarga automatizada del tabulado de viviendas de Morelos para la encuesta intercensal 2015.

# zip::unzip()

Función que sirve para descomprimir zips.

## Ejemplo de uso:

```
zip::unzip("01_Datos/Datos INE/01.zip",  
          exdir = "01_Datos/Datos INE")
```

Descompresión de un zip ubicado en mi carpeta 01\_Datos/Datos INE llamado "01.zip"

## Argumentos importantes:

- \***zipfile**, el nombre y ubicación del archivo \*.zip que se va a descomprimir.
- \***exdir**, Directorio en el cual vamos a colocar los archivos que resulten de la descompresión.

# Función *for*

Función que sirve para crear bucles o ciclos.

## Sintaxis

```
for (elemento in secuencia){  
    Haz_paso_1  
    Haz_paso_2  
    ...  
    Haz_paso_n  
}
```

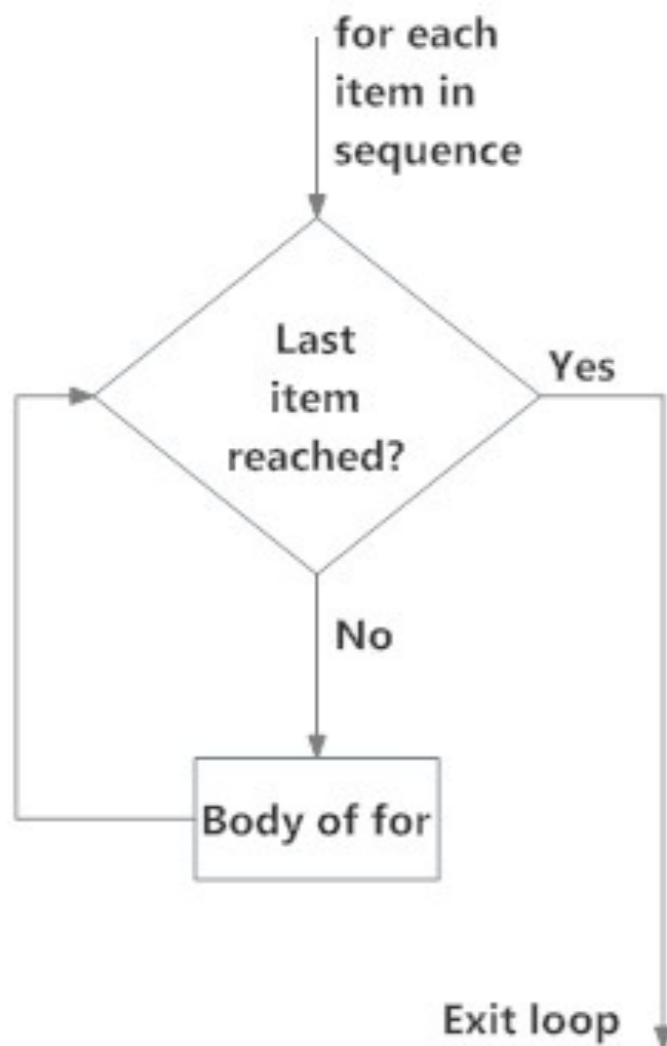


Fig: operation of for loop

# Loops en R con for

Juvenal Campos.

• Quiero hacer un loop:




1. Defino qué quiero hacer...  
... y que resultado esperar  
(Generalmente hacemos loops para automatizar tareas repetitivas).




2. Lo hago para un caso  $n=1$ .  
(Hago ese proceso una vez)

# Loops en R con for

Juvenal Campos.



3. Identifico las partes variables del proceso  
(Qué cambiaría en el caso  $n=2,3,\dots,n$ ?)



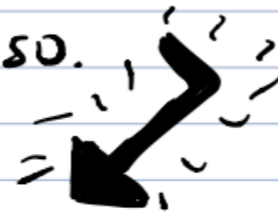
4. Cambio las partes variables por objetos de R.  
(para irlo cambiando en el loop)

# Loops en R con for

Juvenal Campos.



5. Construyo el for  
(o uso algún otro tipo de loop)  
definiendo secuencia, contador  
y proceso.



⇒ ¿algo tronó?

↓ Sí



(Muchas veces las cosas truenan cuando hacemos un loop. Hay que ver que excepción surgió cuando lo ibamos corriendo y ajustar en caso de que sea necesario)

↓ No



Que algo no haya tronado no significa que estemos bien. Hay que verificar si el resultado obtenido es el deseado y, si no, modificar el proceso para corregirlo.

## **Actividad práctica.**

- Repasaremos las funciones vistas previamente y veremos su aplicación en la descarga de datos.**

# Actividad práctica.



## Datos de producción agropecuaria.

Vamos a generar un bucle para descargar los datos del 2015 al 2019 del SIAP (Sistema de Información Agropecuaria y Pesquera) de la Secretaría de Desarrollo Rural.

## Datos de vivienda.

Vamos a descargar los tabulados del cuestionario base para conocer las características de las viviendas en los municipios del país.



# Actividad práctica.



## Datos de Cartografía Electoral.

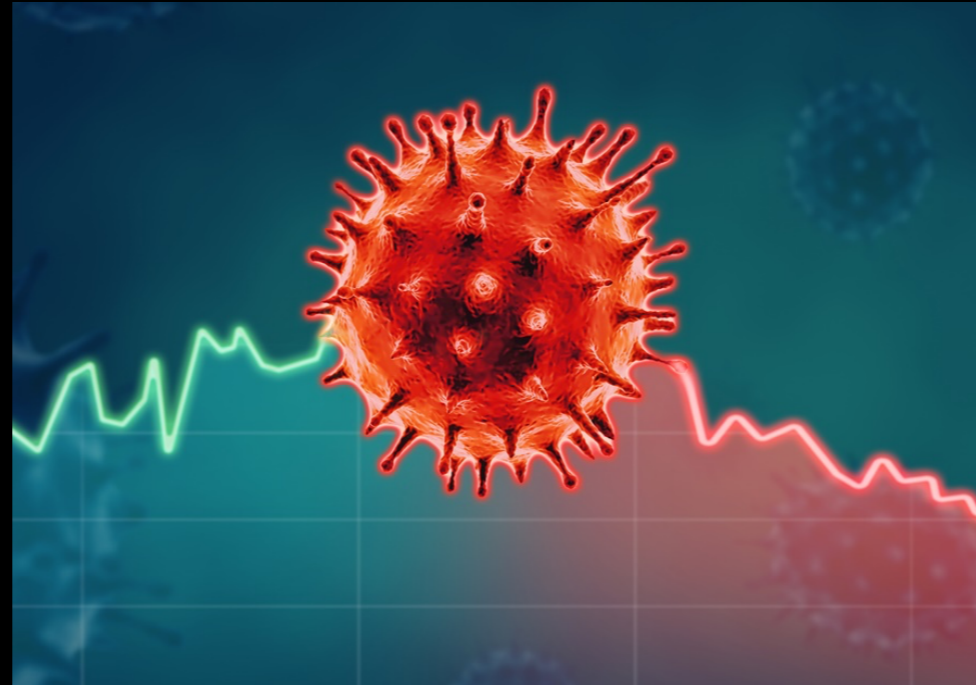
Vamos a descargar los datos de cartografía electoral (los shapefiles) que el INE tiene disponibles.

## Datos de registros diarios de lluvia en la CDMX.

Vamos a descargar los registros diarios de lluvia para las distintas estaciones meteorológicas de la Ciudad de México.



# Actividad práctica.



## **Datos de COVID diarios.**

Vamos a descargar el excel de los datos de casos diarios de COVID publicados el día de ayer (no importa qué día leas esto).