Laboratorio Nacional de Políticas Públicas

CIDE CENTRO DE INVESTIGACIÓN Y DOCENCIA ECONÓMICAS A.C.

Datos Abiertos

(y descarga automatizada)

Visualización y puesta en página web

Septiembre, 2020

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

Datos Abiertos

“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

-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

Datos Abiertos

Beneficios

-Mejorar el e**ntendimiento acerca del buen (o mal) funcionamiento** de las políticas públicas que implementa un gobierno.

-Incrementar la eficiencia gubernamental

-Proveer una **nueva forma de que los ciudadanos se involucren en el trabajo del gobierno,** así como de identificar zonas en las que estos pueden involucrarse, sin depender exclusivamente del gobierno.

-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

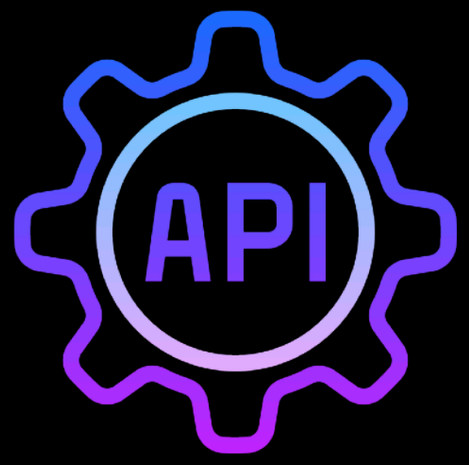
Acceso a los Datos Abiertos

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** (Aplication Program Interface).



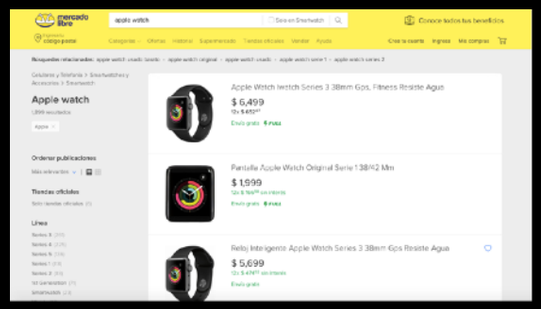
Datos Estructurados y no estructurados

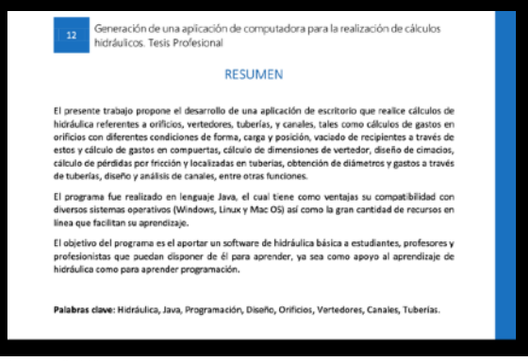
Los datos estructurados son datos presentados en una forma predefinida. Típicamente (pero no exclusivamente) se presenta de forma tabular, en el que las columnas son variables y celdas son valores discretos.

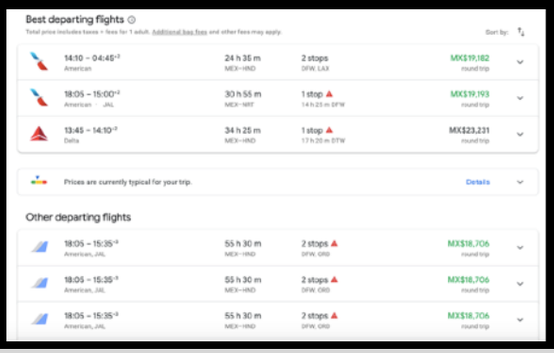
Entre estos datos tenemos los archivos de texto plano (*.csv, *.txt), los archivos de Office (*.docx, *.xlsx), los archivos que almacenan información geográfica (*.shp, *.geojson), entre otros que tienen una estructura definida y son leídos por software de manera nativa.

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.

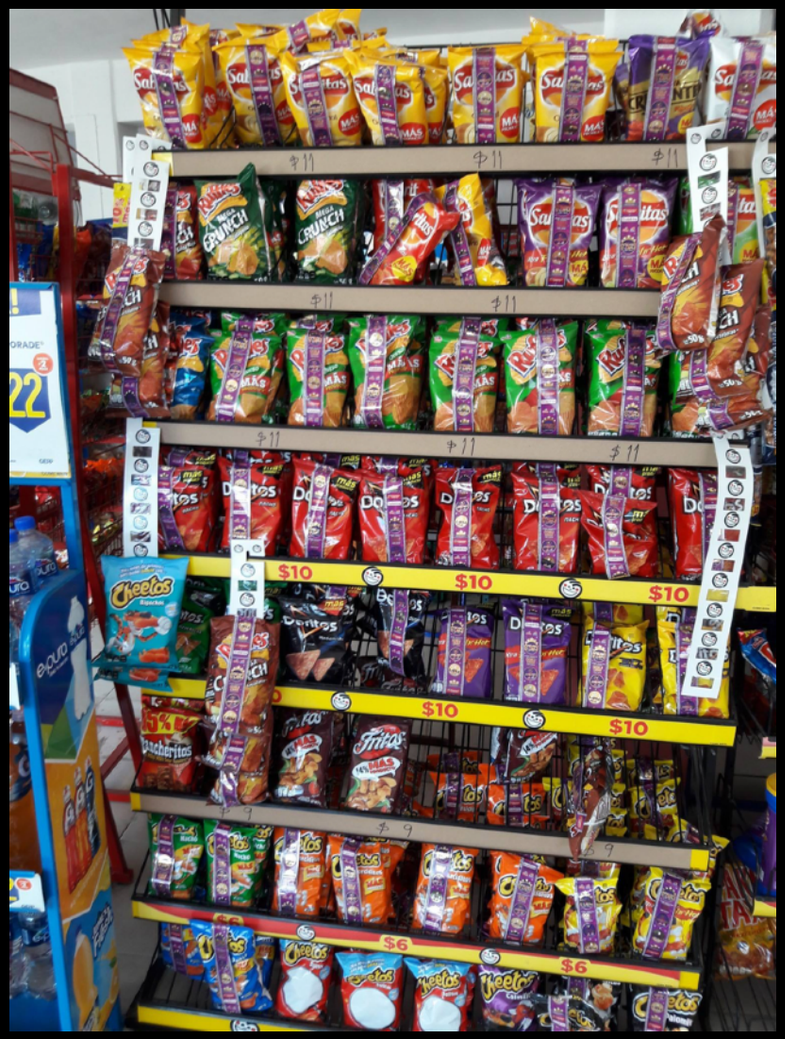
Entre estos se encuentran los *.pdfs (algunos), los *,pdfs fotografías, así como la información disponible en páginas de internet (catálogos, textos, etc).







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 empaquetarlas.

Descarga de datos a través de R y RStudio

Jorge Juvenal Campos Ferreira

Funciones a repasar

- paste() y paste0(), para concatenar texto.

- curl::curl_download(), para descargar archivos de internet (principalmente *.zips y *.xlsx)

- zip::unzip(), para deszipear o descomprimir archivos de internet.

- Función *for* para realizar bucles o ciclos.

paste() y paste0()

Concatenate Strings

Description

Concatenate vectors after converting to character.

Usage

```
paste (..., sep = " ", collapse = NULL)
paste0(..., collapse = NULL)
```

“dos"

Arguments:

1. *sep*, el caracter que separa al texto concatenado.
2. *collapse,* El caracter que pega todos los elementos de un vector. No lo vamos a usar en esta sesión.

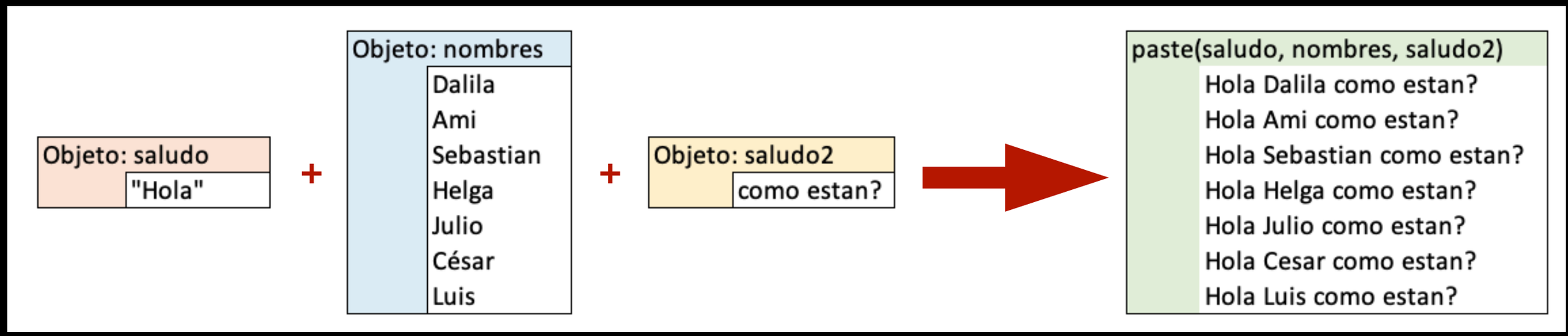
paste() y paste0()

Ejemplos:

```
paste (“Jorge”, “Juvenal”) –> “Jorge Juvenal”
paste0(“Jorge”, “Juvenal”) –> “JorgeJuvenal”
paste0(“Jorge”, “ Juvenal”) –> “Jorge Juvenal”
paste(“Jorge”, “Juvenal”, sep = “___”) –> “Jorge___Juvenal”
paste(c(“uno”, “dos”, “tres”, “cuatro”), collapse = “;”) –>
“uno;dos;tres;cuatro”
```

Aplicado a vectores:

```
paste(saludo, nombres, saludo2)
```



curl::curl_download()

Función que sirve para bajar archivos de internet.

RCurl::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.

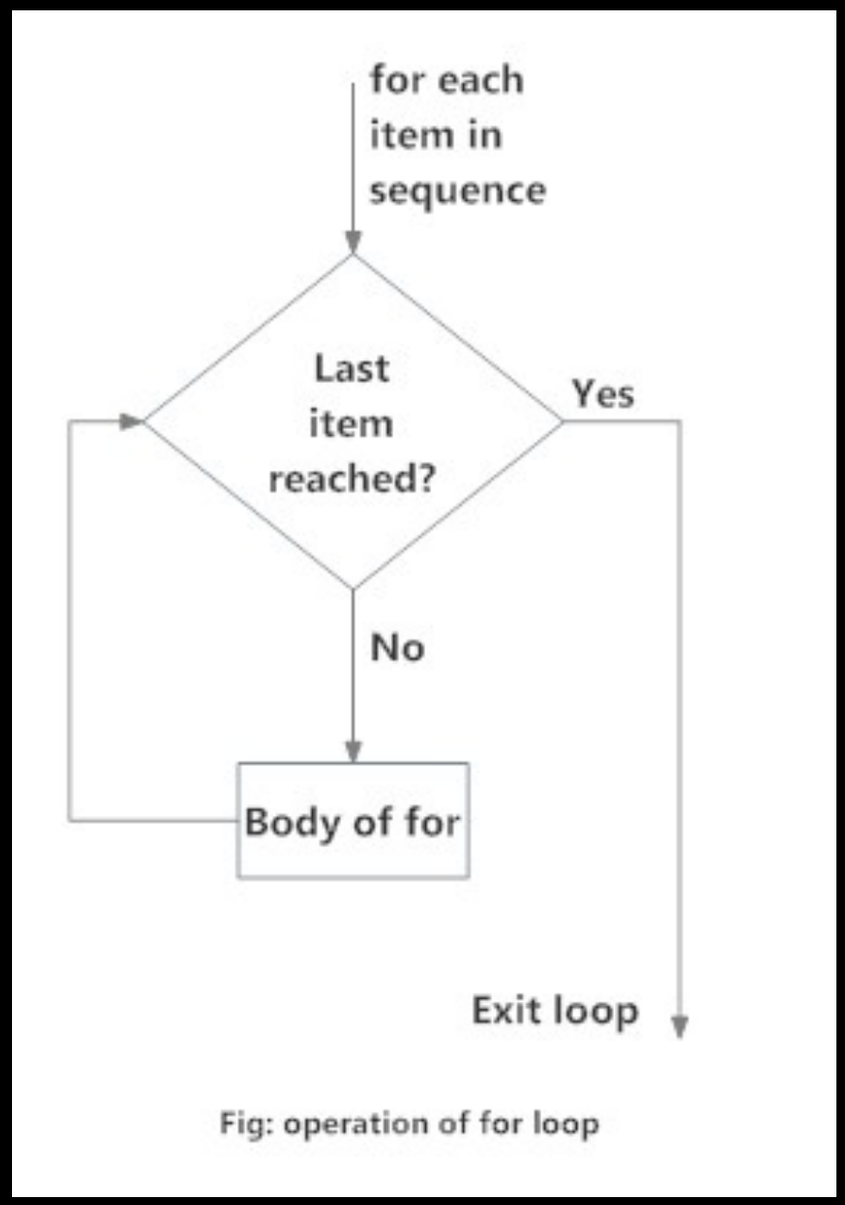


Fig: operation of for loop

Sintaxis

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

Actividad práctica.

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