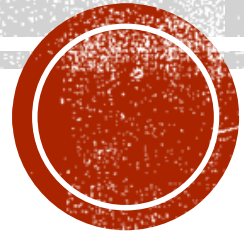
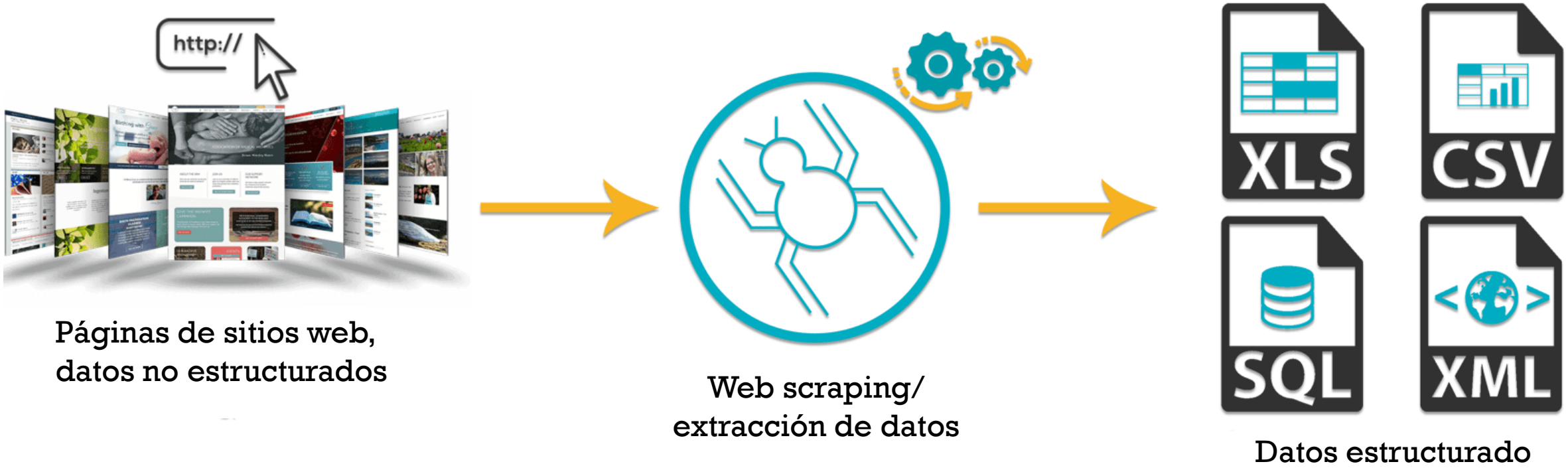
WEB SCRAPING EN R



Por Ángel Sandoval

¿QUÉ ES WEB SCRAPING?

Es un tipo de ”minería de internet” que implica extraer información relevante de un sitio web en particular para su posterior análisis.



Páginas de sitios web, datos no estructurados

Web scraping/ extracción de datos

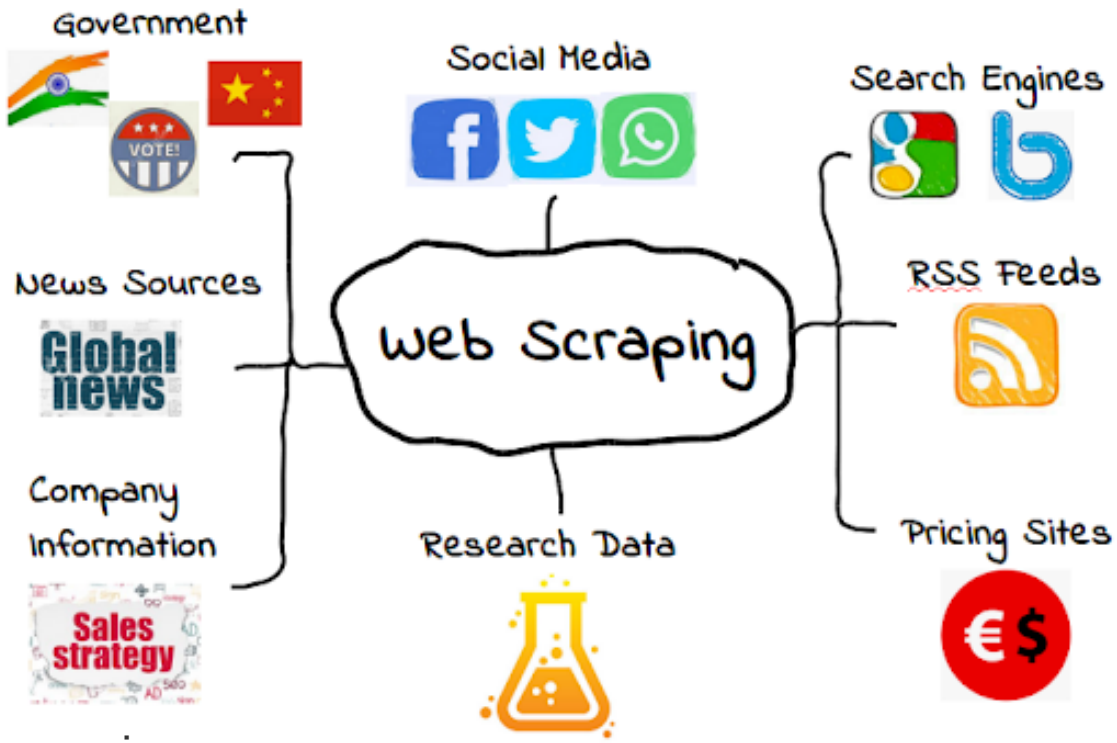
Datos estructurado

Se puede extraer información manualmente pero la idea es automatizar el trabajo utilizando bots

USOS MÁS COMUNES

Extraer datos de contacto como por ejemplo email.

Extraer los títulos y contenidos de un blog.

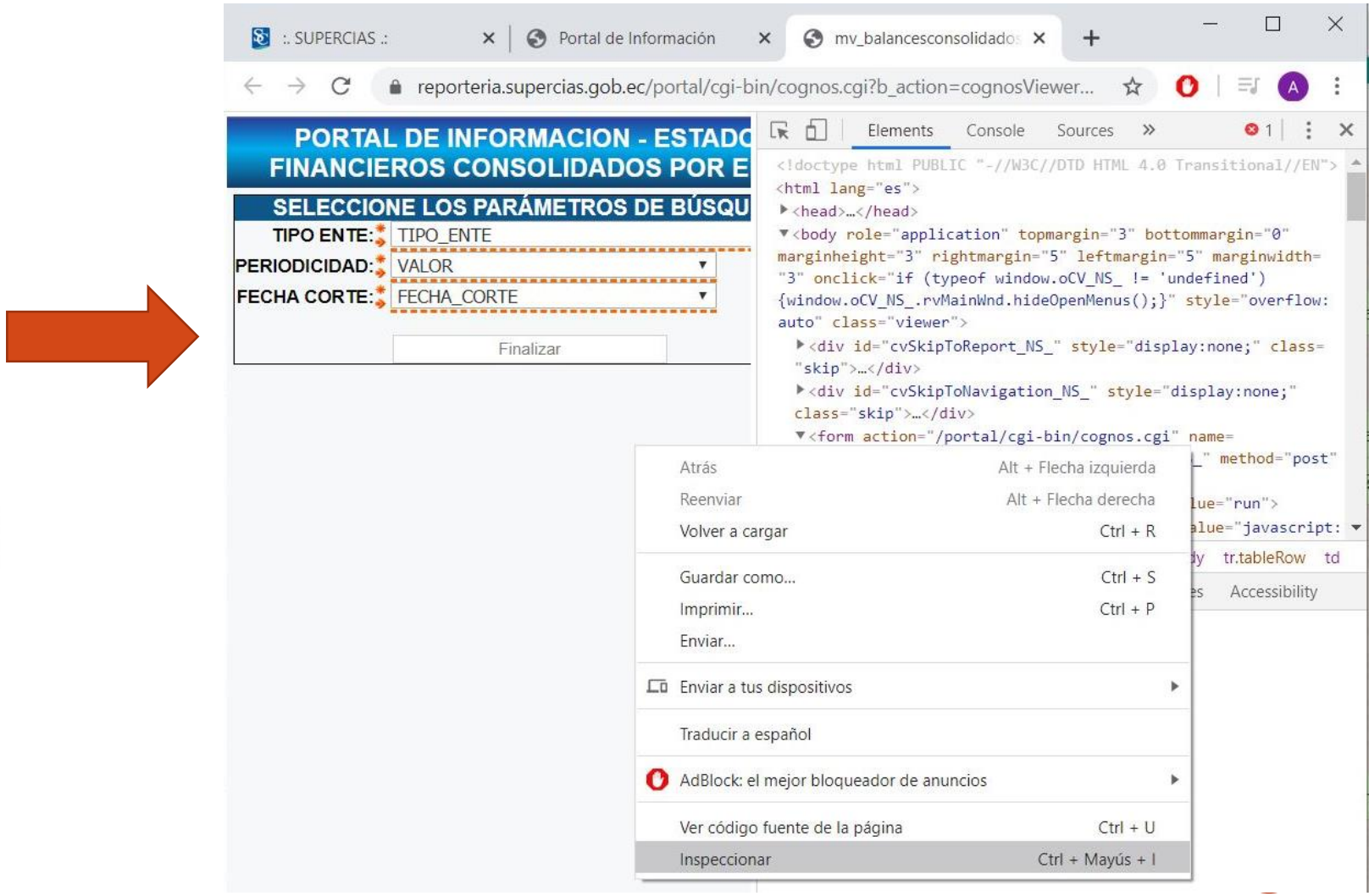


Crear un canal RSS de los contenidos de una página web.

Seguir la evolución de precios de distintos productos.

CONOCIMIENTOS PREVIOS

Estructura básica de una página web



PASO PREVIOS PARA CREAR UN SCRAPIADOR EN R

- 1. Descargar driver para controlar navegador (ejemplo Chrome).
  - https://chromedriver.storage.googleapis.com/2.31/chromedriver_win32.zip
  - https://selenium-release.storage.googleapis.com/3.14/selenium-server-standalone-3.14.0.jar

- 2. Instalar librerías
  - devtools::install_github("johndharrison/binman"). Para gestionar la descarga de archivos binarios
  - devtools::install_github("johndharrison/wdman"). Proporciona funciones para descargar estos archivos binarios y para gestionar los procesos que los involucran
  - devtools::install_github("ropensci/RSelenium")

COMO HACER QUE FUNCIONE LA HERRAMIENTA

- 1. En el CMD setear la dirección donde se guardo los driver

- 2. Ejecutar el driver especificando un puerto
  - java -jar selenium-server-standalone-3.14.0.jar -port 8002.

La sentencia anterior debe arrojar como resultado en el cmd:
Selenium Server us up and running port 8002