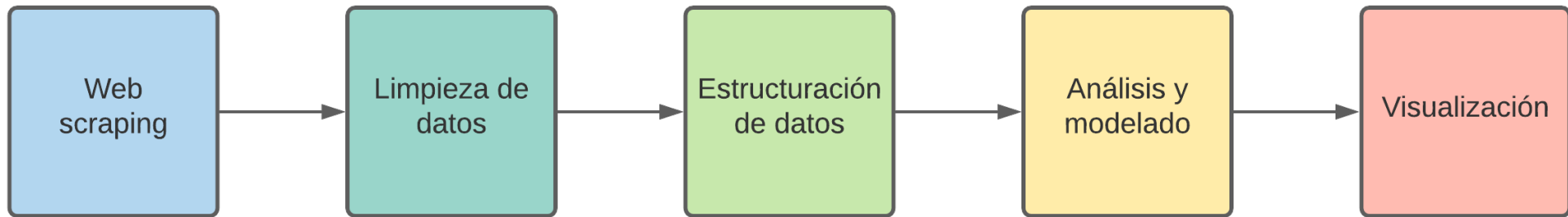


# Ejemplos prácticos de web scraping dinámica y otros usos del manejo de texto

Javier Mtz

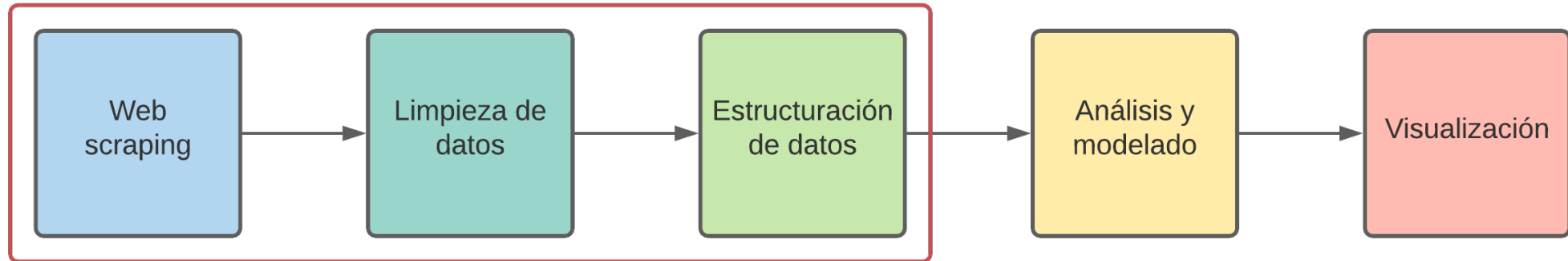
# ¿Qué vamos a ver?

## Proceso de análisis



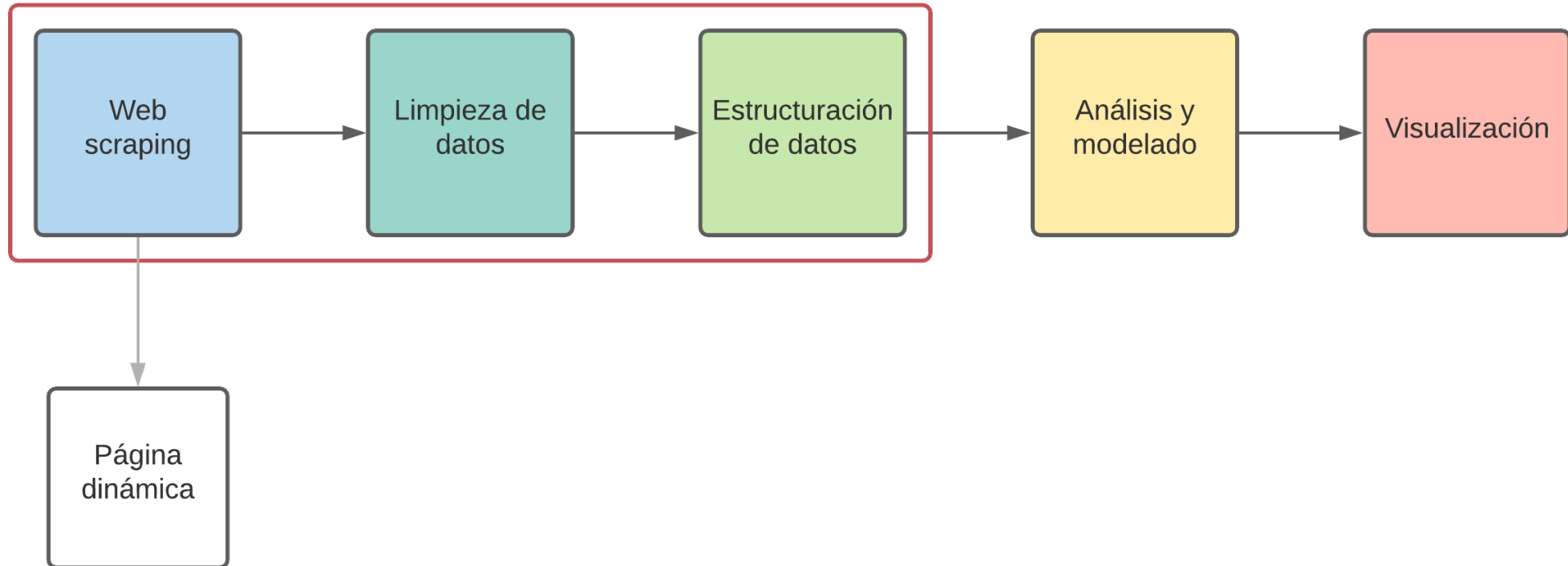
# ¿Qué vamos a ver?

## Proceso de análisis



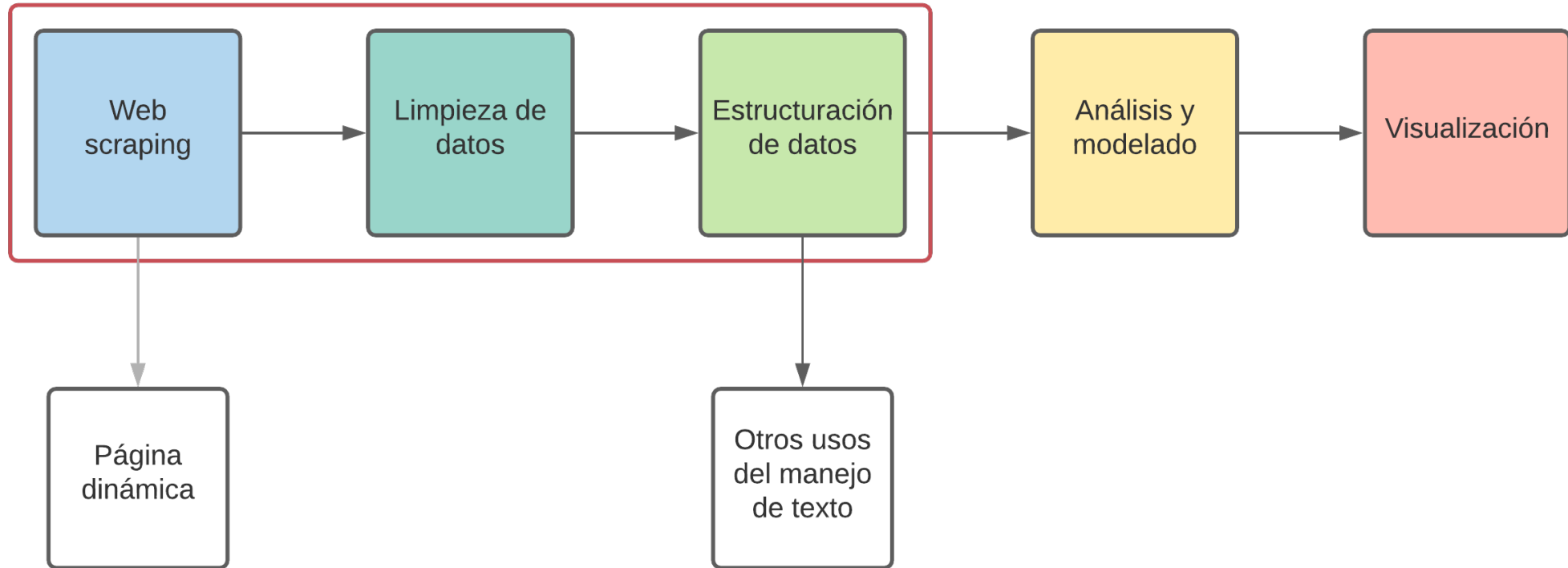
# ¿Qué vamos a ver?

## Proceso de análisis



# ¿Qué vamos a ver?

## Proceso de análisis



# Análisis de documentos de CompraNet

# Análisis de documentos de CompraNet

## Objetivo:

- Ampliar la información disponible de contrataciones públicas.

# Análisis de documentos de CompraNet

Objetivo:

Yo intentando hacer un análisis con datos públicos:





# Análisis de documentos de CompraNet

## Objetivo:

- Ampliar la información disponible de contrataciones públicas.
- Mapear información documentos e información disponible de CompraNet.
- Análisis de texto de documentos relevantes.

# Análisis de documentos de CompraNet

## Identificación de fuentes de información

- compranet.hacienda.gob.mx

00:24 CST - Central America Time DST

**CompraNet** **SHCP** SECRETARÍA DE HACIENDA

Anuncios Vigentes    Anuncios en seguimiento o concluidos

[Página de Inicio](#)

Introduzca Filtro (escriba para iniciar la)

	Nombre de la Unidad Compradora (UC)	Referencia del Expediente	Descripción del Expediente	Tipo de Contratación	Plazo de participación o vigencia del anuncio
1	IMSS-Unidad Médica de Alta Especialidad, Hospital de GinecoObstetricia del Centro Médico Nacional de Occidente #050GYR079	AA-050GYR079-E246-2021	CONTRATACIÓN DE MATERIAL DE ASEO FOCON 155	Adquisiciones	26/08/2021 08:00
2	IMSS-Unidad Médica de Alta Especialidad, Hospital de GinecoObstetricia del Centro Médico Nacional de Occidente #050GYR079	AA-050GYR079-E247-2021	SEGUNDA VUELTA DE CONTRATACIÓN DE MEDICAMENTOS FOCON 151	Adquisiciones	26/08/2021 08:00
3	IMSS-Coordinación de Abastecimiento y Equipamiento #050GYR014	AA-050GYR014-E636-2021	AA-050GYR014-E636-2021 MEDICAMENTOS COVID-19	Adquisiciones	26/08/2021 08:00
4	IMSS-Unidad Médica de Alta Especialidad, Hospital de GinecoObstetricia del Centro Médico Nacional de Occidente #050GYR079	AA-050GYR079-E249-2021	SEGUNDA VUELTA DE CONTRATACIÓN MATERIAL DE CURACIÓN FOCON 150	Adquisiciones	26/08/2021 08:00
5	ISSSTE-Departamento de Recursos Materiales y Obras #051GYN060	IO-051GYN060-E26-2021	Rehabilitación integral de área de urgencias	Servicios Relacionados con la OP	26/08/2021 08:00
6	SEDENA-Subdireccion de Adquisiciones #007000999	LA-007000999-E721-2021	Adqs. Egmo. H.M.R. Torreón, Coah. y H.M.Z. Ixcotel, Oax.	Adquisiciones	26/08/2021 08:00

ADO. RFFACS. MANTO PREVENTIVO Y

- Expedientes de CompraNet

**CompraNet** **SHCP** SECRETARÍA DE HACIENDA

[Volver a la Lista](#)    [Ingresar al sistema CompraNet](#)

Expediente 2317227 - CONTRATACIÓN DE MATERIAL DE ASEO FOCON 155  
Referencia del Expediente AA-050GYR079-E246-2021

**Detalles del Expediente**

Anuncio Publicado

**Detalles del Expediente**

<b>Código del Expediente</b> 2317227	<b>Descripción del Expediente</b> CONTRATACIÓN DE MATERIAL DE ASEO FOCON 155
<b>Referencia del Expediente</b> AA-050GYR079-E246-2021	<b>Tipo de Expediente</b> 05. Adjudicación Directa LAASP
<b>Categorías del Expediente</b> 2150-Material de limpieza	

**Detalles del Anuncio**

<b>Descripción del Anuncio</b> CONTRATACIÓN DE MATERIAL DE ASEO FOCON 155	<b>Notas</b> Notas Adicionales por Defecto
<b>Tipo de Contratación</b> Adquisiciones	<b>Entidad Federativa</b> Jalisco
<b>Fecha de publicación del anuncio (Convocatoria / Invitación / Adjudicación / Proyecto de Convocatoria)</b> 20/08/2021 13:52	<b>Plazo de participación o vigencia del anuncio</b> 26/08/2021 08:00
<b>Fecha de Inicio del Proceso</b>	<b>Fecha de Cierre del Proceso</b>

# Análisis de documentos de CompraNet

¿Qué queremos extraer de la página?

00:24 CST - Central America Time DST

CompraNet

SHCP  
SECRETARÍA DE HACIENDA

Anuncios VigentesAnuncios en seguimiento o concluidos

Página de Inicio

Introduzca Filtro (escriba para iniciar la)


	Nombre de la Unidad Compradora (UC)	Referencia del Expediente	Descripción del Expediente	Tipo de Contratación	Plazo de participación o vigencia del anuncio
1	IMSS-Unidad Médica de Alta Especialidad, Hospital de GinecoObstetricia del Centro Médico Nacional de Occidente #050GYR079	AA-050GYR079-E246-2021	CONTRATACIÓN DE MATERIAL DE ASEO FOCON 155	Adquisiciones	26/08/2021 08:00
2	IMSS-Unidad Médica de Alta Especialidad, Hospital de GinecoObstetricia del Centro Médico Nacional de Occidente #050GYR079	AA-050GYR079-E247-2021	SEGUNDA VUELTA DE CONTRATACIÓN DE MEDICAMENTOS FOCON 151	Adquisiciones	26/08/2021 08:00
3	IMSS-Coordinación de Abastecimiento y Equipamiento #050GYR014	AA-050GYR014-E636-2021	AA-050GYR014-E636-2021 MEDICAMENTOS COVID-19	Adquisiciones	26/08/2021 08:00
4	IMSS-Unidad Médica de Alta Especialidad, Hospital de GinecoObstetricia del Centro Médico Nacional de Occidente #050GYR079	AA-050GYR079-E249-2021	SEGUNDA VUELTA DE CONTRATACIÓN MATERIAL DE CURACIÓN FOCON 150	Adquisiciones	26/08/2021 08:00
5	ISSSTE-Departamento de Recursos Materiales y Obras #051GYN060	IO-051GYN060-E26-2021	Rehabilitación integral de área de urgencias	Servicios Relacionados con la OP	26/08/2021 08:00
6	SEDENA-Subdireccion de Adquisiciones #007000999	LA-007000999-E721-2021	Adqs. Egmt. H.M.R. Torreón, Coah. y H.M.Z. Ixcotel, Oax.	Adquisiciones	26/08/2021 08:00
			ADD. RFFACS: MANTO PREVENTIVO Y		

# Análisis de documentos de CompraNet

¿Qué queremos extraer de la página?

- Registros y links de expedientes

00:24 CST - Central America Time DST

**CompraNet** 

Anuncios Vigentes    Anuncios en seguimiento o concluidos

[Página de Inicio](#)

Introduzca Filtro (escriba para iniciar la)

	Nombre de la Unidad Compradora (UC)	Referencia del Expediente	Descripción del Expediente	Tipo de Contratación	Plazo de participación o vigencia del anuncio
1	IMSS-Unidad Médica de Alta Especialidad, Hospital de GinecoObstetricia del Centro Médico Nacional de Occidente #050GYR079	AA-050GYR079-E246-2021	CONTRATACIÓN DE MATERIAL DE ASEO FOCON 155	Adquisiciones	26/08/2021 08:00
2	IMSS-Unidad Médica de Alta Especialidad, Hospital de GinecoObstetricia del Centro Médico Nacional de Occidente #050GYR079	AA-050GYR079-E247-2021	SEGUNDA VUELTA DE CONTRATACIÓN DE MEDICAMENTOS FOCON 151	Adquisiciones	26/08/2021 08:00
3	IMSS-Coordinación de Abastecimiento y Equipamiento #050GYR014	AA-050GYR014-E636-2021	AA-050GYR014-E636-2021 MEDICAMENTOS COVID-19	Adquisiciones	26/08/2021 08:00
4	IMSS-Unidad Médica de Alta Especialidad, Hospital de GinecoObstetricia del Centro Médico Nacional de Occidente #050GYR079	AA-050GYR079-E249-2021	SEGUNDA VUELTA DE CONTRATACIÓN MATERIAL DE CURACIÓN FOCON 150	Adquisiciones	26/08/2021 08:00
5	ISSSTE-Departamento de Recursos Materiales y Obras #051GYN060	IO-051GYN060-E26-2021	Rehabilitación integral de área de urgencias	Servicios Relacionados con la OP	26/08/2021 08:00
6	SEDENA-Subdireccion de Adquisiciones #007000999	LA-007000999-E721-2021	Adqs. Egmt. H.M.R. Torreón, Coah. y H.M.Z. Ixcotel, Oax.	Adquisiciones	26/08/2021 08:00

ADO. RFFACS. MANTO PREVENTIVO Y

# Análisis de documentos de CompraNet

¿Qué queremos extraer de la página?

- Registros y links de expedientes
- Detalles del expediente

The screenshot displays the CompraNet interface. At the top, there's a navigation bar with the 'CompraNet' logo and the 'SHCP SECRETARÍA DE HACIENDA' logo. Below the navigation bar, there's a header section with a 'Volver a la Lista' link and an 'Ingresar al sistema CompraNet' button. The main content area shows the details for 'Expediente 2317227 - CONTRATACIÓN DE MATERIAL DE ASEO FOCON 155' with the reference 'AA-050GYR079-E246-2021'. A 'Detalles del Expediente' tab is selected, showing a table with the following data:

Código del Expediente	Descripción del Expediente
2317227	CONTRATACIÓN DE MATERIAL DE ASEO FOCON 155
Referencia del Expediente	Tipo de Expediente
AA-050GYR079-E246-2021	05. Adjudicación Directa LAASP
Categorías del Expediente	
2160-Material de limpieza	

Below the table, there's a 'Detalles del Anuncio' section with the following data:

Descripción del Anuncio	Notas
CONTRATACIÓN DE MATERIAL DE ASEO FOCON 155	Notas Adicionales por Defecto
Tipo de Contratación	Entidad Federativa
Adquisiciones	Jalisco
Fecha de publicación del anuncio (Convocatoria / Invitación / Adjudicación / Proyecto de Convocatoria)	Plazo de participación o vigencia del anuncio
20/08/2021 13:52	26/08/2021 08:00
Fecha de inicio del Contrato	Formación del Contrato

# Análisis de documentos de CompraNet

¿Qué queremos extraer de la página?

- Registros y links de expedientes
- Detalles del expediente
- Datos generales y anexos

Expediente 2317089 - Modelación hidráulica de una línea de conducción y de la calidad del agua con él

[Volver a la Lista](#) [Ingresar al sistema CompraNet](#)

DATOS GENERALES DEL PROCEDIMIENTO DE CONTRATACIÓN		
👁	Número del Procedimiento (Expediente)	• Este número se generará al momento de publicar el Procedimiento. AA-016RJE001-E132-2021
👁	Carácter del procedimiento	• Indicar el carácter del procedimiento Nacional
👁	Medio o forma del procedimiento	• Seleccionar el medio o forma de participación en el procedimiento. (Electrónica sólo aplica a la LAASPP) Electrónica
👁	Procedimiento exclusivo para MIPYMES	• Define si se establece como requisito de participación el que los licitantes acrediten ser una micro, pequeña o mediana empresa (sólo aplica a la LAASPP) No

ANEXOS DEL PROCEDIMIENTO DE CONTRATACIÓN		
👁	Datos relevantes de contrato	Archivo que contiene el informe con los datos relevantes del contrato. Tamaño máximo 150 MB. (sin archivo adjunto)
👁	Escrito de justificación de la excepción a la licitación pública, cuando aplique	Archivo que contiene el escrito de justificación de la excepción a la licitación pública fundada en el artículo 41 de la LAASPP o en el artículo 42 de la LOPSRM (sin archivo adjunto)
👁	Testimonio del Testigo Social, si el procedimiento contó con el	Archivo que contiene el testimonio del testigo social. Tamaño máximo 150 MB. (sin archivo adjunto)
👁	Documento de autorización de subcontratación	Sólo aplica a la LOPSRM, en los casos en que haya existido alguna subcontratación no prevista en la convocatoria, invitación o condiciones de adjudicación, adjunte el o los documentos de autorización emitidos por el área responsable de la ejecución de los trabajos (sin archivo adjunto)

Procedimiento

# Análisis de documentos de CompraNet

¿Qué queremos extraer de la página?

- Registros y links de expedientes
- Detalles del expediente
- Datos generales y anexos
- Anexos adicionales

Expediente 2053203 - AA-0550YR031-E29-2020 ADQ. DE MEDICAMENTO, MAT DE CURACIÓN Y LABORATORIO (RVP)

Página Principal [Ingresar al sistema CompraNet](#)

Nombre del archivo	Descripción del archivo	Comentarios	Última fecha de modificación
1 <a href="#">AA-0550YR031-E29-2020.doc (456 KB)</a>	CONVOCATORIA		27/01/2020 12:10
2 <a href="#">DRC DDP0151.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		06/02/2020 10:52
3 <a href="#">DRC DDP0152.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
4 <a href="#">DRC DDP0153.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
5 <a href="#">DRC DDP0154.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
6 <a href="#">DRC DDP0155.pdf (101 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
7 <a href="#">DRC DDP0156.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
8 <a href="#">DRC DDP0157.pdf (101 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
9 <a href="#">DRC DDP0158.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
10 <a href="#">DRC DDP0159.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
11 <a href="#">DRC DDP0160.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
12 <a href="#">DRC DDP0161.pdf (101 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
13 <a href="#">DRC DDP0162.pdf (99 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
14 <a href="#">DRC DDP0163.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
15 <a href="#">DRC DDP0164.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
16 <a href="#">DRC DDP0165.pdf (101 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
17 <a href="#">DRC DDP0166.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
18 <a href="#">DRC DDP0167.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
19 <a href="#">DRC DDP0168.pdf (98 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29
20 <a href="#">DRC DDP0169.pdf (99 KB)</a>	DATOS RELEVANTES DE CONTRATO		12/02/2020 12:29

Anexos: 114



# Análisis de documentos de CompraNet

Dificultades:

1. Extraer información de una página dinámica
2. Establecer procedimiento de consulta





# Análisis de documentos de CompraNet

## Extraer información de una página dinámica

- A diferencia de las páginas estáticas, éstas reaccionan ante la información y acciones que realicemos.
- Estas páginas necesitan de nuestras *respuestas* para funcionar. Por tanto, cada visitante puede ver diferente información.
- Muchas de estas páginas generan información posteriormente a que cargó.



# Análisis de documentos de CompraNet

Extraer información de una página dinámica y procedimiento de consulta

¿Qué pasa con CompraNet si tratamos de analizar una página como si fuera estática?

```
library(rvest)

read_html("https://compranet.hacienda.gob.mx/esop/toolkit/opportunity/ns/1795310/detail.si?
isOnModification=false&_ncp=1629966548970.111568-2") %>%
  html_table()
```

```
## Error in open.connection(x, "rb"): HTTP error 401.
```

**HTTP 401:** indica que la petición (*request*) no ha sido ejecutada porque carece de credenciales válidas de autenticación para el recurso solicitado.



# Análisis de documentos de CompraNet

Extraer información de una página dinámica y procedimiento de consulta

¿Qué podemos hacer al respecto?

Usar `{RSeelenium}` en tres pasos:

1. Crear el procedimiento de análisis de la página dinámica.
2. Obtener el código fuente en R.
3. Seleccione la información exacta necesaria del código fuente.



# Análisis de documentos de CompraNet

Extraer información de una página dinámica y procedimiento de consulta

¿Qué podemos hacer al respecto?

Usar `{RSeelenium}`

- Acceso a varios recursos.
- Selenium utiliza métodos y no sólo funciones.
- Permite interactuar con el navegador y los elementos de la página al mismo tiempo.



# Análisis de documentos de CompraNet

¿Cómo vamos a mapear la información de los expedientes?

1. Preparar navegador
2. Establecer procedimiento para consultar cada expediente
3. Extraer la información del expediente seleccionado
4. Repetir procedimiento para demas expedientes
5. Descargar y procesar documentos



# Análisis de documentos de CompraNet

## 1. Preparar navegador

- Prepara contenedor de Dockers para evitar problemas de Selenium
- Qué es y por qué usar Dockers
- Seleccionar *drivers* necesarios



# Análisis de documentos de CompraNet

## 1. Preparar navegador

Establecer perfil

```
firefox_profile.me <- makeFirefoxProfile(list(browser.download.dir = normalizePath([carpeta_contratos]),  
                                             browser.helperApps.neverAsk.openFile = [tipos de documentos],  
                                             browser.helperApps.neverAsk.saveToDisk = [tipos de documentos]))
```



# Análisis de documentos de CompraNet

## 1. Preparar navegador

Crear y lanzar driver remoto

- `rsDriver` inicia un server y a partir de él podemos utilizar el navegador.
- Cuidar que el *port* coincida con el contenedor y no iniciar dos servers en el mismo.

```
driver <- rsDriver(browser = "firefox",  
                  port = 491L,  
                  extraCapabilities = firefox_profile.me)
```

```
browser <- driver[["client"]]  
server <- driver[["server"]]
```





# Análisis de documentos de CompraNet

## 1. Preparar navegador

- Como resultado, se abrirá un navegador (este no será observable si lo creamos en un contenedor de Dockers)
- Se creará un objeto llamado *driver*.



# Análisis de documentos de CompraNet

## 1. Preparar navegador

Para utilizar el navegador podemos aplicarle métodos.

```
browser$metodo()
```

Algunos métodos:

- `browser$navigate("https://pagina.com")`: sirve para abrir la página establecida.
- `browser$goBack()`: regresar a la página anterior.
- `browser$goForward()`: ir a la siguiente página.



# Análisis de documentos de CompraNet

## 1. Preparar navegador

Algunos métodos:

- `browser$refresh()`: volver a cargar página.
- `browser$CurrentUrl()`: obtener la url de la página.
- `browser$close()`: cerrar navegador.
- `browser$getPageSource()[[1]]`: obtener la página fuente.
  - Este método regresa una lista. El primer elemento es el xml de la página `[[1]]`



# Análisis de documentos de CompraNet

## 1. Preparar navegador

Algunos métodos:

- `browser$findElement(using = selector, value)`: encontrar ubicación de elemento.
  - El selector puede ser xpath, css, id, name, link text.

```
icono <- browser$findElement(using = selector, value)
```

- `icono$clickElement()`: este método permite hacer click sobre el elemento previamente seleccionado.



# Análisis de documentos de CompraNet

## 2. Establecer procedimiento para consultar expediente

```
browser$navigate("https://compranet.hacienda.gob.mx/esop/guest/go/public/opportunity/past?locale=es_MX")
pagina_actual <- read_html(browser$pageSource()[[1]])
tabla_expedientes <- pagina_actual %>%
  html_table()

exp <- browser$findElement(using = "partial link text",
                           tabla_expedientes[1,4])

exp$clickElement()
```



# Análisis de documentos de CompraNet

## 2. Establecer procedimiento para consultar expediente

tabla\_expedientes

Número de Fila	Columna de Icono	Nombre de la Unidad Compradora (UC)	Referencia del Expediente	Descripción del Expediente	Tipo de Contratación	Plazo de participación o vigencia del anuncio
1	NA	MICH-Arteaga-Dirección de Programación y Presupuesto #816010744	HAAM-CONTRATO-FORFIN-06-2017	CONSTRUCCION DE COLECTOR MARGINAL DE AGUAS NEGRAS Y LAGUNAS FACULTATIVAS, ANAERO	Obra Pública	31/08/2021 00:00
2	NA	SEDENA-Subdireccion de Adquisiciones #007000999	LA-007000999-E840-2021	SV. MANTO.AEROGENERADORES PARQUE EOLICO SDN 2/a. VUELTA	Servicios	31/08/2021 00:00
3	NA	MEX-Universidad Intercultural del Estado de México-Dirección de Administración y Finanzas #915110916	210C2601080003L/ICTP/002/2021	ADQUISICION DE MATERIALES Y UTILES DE OFICINA	Adquisiciones	31/08/2021 00:00
4	NA	JAL-Secretaría de Salud-Servicios de Salud Jalisco #914010985	IA-914010985-E34-2021	MEDICAMENTOS PARA LAS UNIDADES MÉDICAS DEL O.P.D SERVICIOS DE SALUD JALISCO	Adquisiciones	31/08/2021 00:00

# Análisis de documentos de CompraNet

## 3. Extraer la información del expediente seleccionado

```
expediente_actual <- read_html(browser$getPageSource()[[1]])

tablas_exp <- expediente_actual %>%
  html_table()

texto <- expediente_actual %>%
  html_nodes(".form_container:nth-child(9) ul") %>%
  html_text() %>%
  str_remove_all("\t")

datos_exp <- texto %>%
  str_split("\n") %>%
  .[[1]] %>%
  .[str_detect(., "")]

datos_exp <- tibble(categoria = datos_exp[!((1:length(datos_exp) %% 2) == 0)],
  informacion = datos_exp[(1:length(datos_exp) %% 2) == 0])

datos_exp
```



# Análisis de documentos de CompraNet

## 3. Extraer la información del expediente seleccionado

categoria	informacion
Código del Expediente	2053203
Descripción del Expediente	AA-050GYR031-E29-2020 ADQ. DE MEDICAMENTO, MAT DE CURACIÓN Y LABORATORIO (RVP)



# Análisis de documentos de CompraNet

## 3. Extraer la información del expediente seleccionado

```
scrap_docs <- map_df(tablas_exp,
  function(x){
    x %>% clean_names() %>%
    {if(any(str_detect(names(.), "datos_generales_del_procedimiento"))){
      rename(., datos_generales = contains("datos_generales")) %>%
      mutate(comienza_anexo = str_detect(str_to_upper(datos_generales1), "ANEXOS"),
        comienza_anexo_num = cumsum(comienza_anexo)) %>%
      filter(comienza_anexo_num > 0) %>% filter(row_number() > 2) %>%
      transmute(categoria = datos_generales2,nombre_archivo = str_replace_all(datos_generales4, "\\)
[^_]*", "\\)"))} else .} %>%
    {if(any(str_detect(names(.), "nombre_del_archivo"))){
      rename(., nombre_archivo = contains("nombre"),
        categoria = contains("descripcion")) %>%
      select(nombre_archivo, categoria) } else .} %>%
    add_column(., !!!c(nombre_archivo = NA_real_, categoria = NA_real_)[setdiff(names(c(nombre_archivo
= NA_real_, categoria = NA_real_)), names(.))]) %>%
    mutate(nombre_archivo = as.character(nombre_archivo), categoria = as.character(categoria)) %>%
    select(nombre_archivo, categoria) %>%
    filter(!is.na(nombre_archivo) & !is.na(categoria))
  })
```

scrap\_docs



# Análisis de documentos de CompraNet

## 3. Extraer la información del expediente seleccionado

nombre_archivo	categoria
(sin archivo adjunto)	Datos relevantes de contrato
DP-050GYR031-E29-2020.pdf (44 KB)	Escrito de justificación de la excepción a la licitación pública, cuando aplique
(sin archivo adjunto)	Testimonio del Testigo Social, si el procedimiento contó con el
(sin archivo adjunto)	Documento de autorización de subcontratación
AA-050GYR031-E29-2020.doc (456 KB)	CONVOCATORIA

# Análisis de documentos de CompraNet

## 4. Repetir procedimiento para demas expedientes

```
extraccion_expediente <- function(referencia_del_expediente) {  
  # Entrar a la página del expediente  
  # Extraer información del expediente  
  # Crear dataframe de documentos  
  # Cruzar información en un dataframe  
}  
  
for (ref_exp in tabla_expedientes$`Referencia del Expediente`) {  
  extraccion_expediente(ref_exp)  
}  
  
new_page <- try(browser$findElement(using = "class", ".NavBtnForward"),  
               silent = T)  
  
while (new_page != "try-error") {  
  for (ref_exp in tabla_expedientes$`Referencia del Expediente`) {  
    extraccion_expediente(ref_exp)  
  }  
  
  new_page <- try(browser$findElement(using = "class", ".NavBtnForward"),  
                 silent = T)  
}
```



# Análisis de documentos de CompraNet

## 4. Repetir procedimiento para demas expedientes

num_exp	url_exp	nombre_archivo	categoria
2161649	https://compranet.hacienda.gob.mx/esop/guest/go/opportunity/detail?opportunityId=1891257	TM-TRAMO4 20-OI-04 DR.pdf (143 KB)	Datos relevantes de contrato
2161649	https://compranet.hacienda.gob.mx/esop/guest/go/opportunity/detail?opportunityId=1891257	(sin archivo adjunto)	Escrito de justificación de la excepción a la licitación pública, cuando aplique
2161649	https://compranet.hacienda.gob.mx/esop/guest/go/opportunity/detail?opportunityId=1891257	Testimonio Final AO-021W3N003-E229-2020.... (833 KB)	Testimonio del Testigo Social, si el procedimiento contó con el
2161649	https://compranet.hacienda.gob.mx/esop/guest/go/opportunity/detail?opportunityId=1891257	(sin archivo adjunto)	Documento de autorización de subcontratación
2161649	https://compranet.hacienda.gob.mx/esop/guest/go/opportunity/detail?opportunityId=1891257	Testimonio Tramo 4.pdf (833 KB)	TESTIMONIO TESTIGO SOCIAL

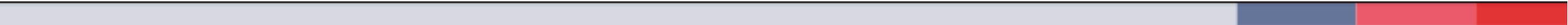
# Análisis de documentos de CompraNet

## 5. Descargar y procesar documentos

```
extraccion_expediente <- function(referencia_del_expediente) {  
  # Entrar a la página del expediente  
  # Extraer información del expediente  
  # Crear dataframe de documentos  
  # Cruzar información en un dataframe  
  
  docs_descarga <- scrap_docs$nombre_archivo %>%  
    .[!str_detect(., "(sin archivo adjunto)")]  
  
  for (doc in docs_descarga) {  
    ub_doc <- try(driver$findElement(using = "partial link text", doc),  
                 silent = T)  
  
    ub_doc$clickElement()  
  
    Sys.sleep(5)  
  }  
}  
  
pdftools::pdf_text(pdf = [Ubicacion])
```



Gracias



---