Reporte Técnico Extracción de Datos

Secretaría de Seguridad y Protección Ciudadana (SSPC)

Mariana Consuelo Fernández Espinosa
Yadira Elizabeth Peralta Torres
Fernando Alarid Escudero
Regina Isabel Medina Rosales

PADeCI

Septiembre de 2020

Introducción

Comúnmente se menciona que el 80 % del trabajo relacionado al análisis de datos corresponde al proceso de limpieza y tratamiento de los datos (Referencia) es por esto que se han desarrollado a lo largo del tiempo múltiples metodologías que estandarizan los métodos de procesamiento de datos y otorgan un marco definido para identificar cada paso en el tratamiento de la información, sin embargo cada investigación posee características únicas donde apegarse a formas definidas de trabajo puede resultar complicado e ineficiente para la investigación por tal motivo es apremiante otorgar atención a la importancia del desarrollo del procesamiento de los datos desde la etapa inicial a la etapa final donde se entregan los resultados ordenados para su posterior análisis.

El siguiente reporte muestra de forma detallada el proceso que se siguió para extraer, limpiar y organizar los datos provenientes de la fuente (Informe de seguridad). Cada paso describe la metodología seleccionada y adecuada a los fines de la investigación, con el objetivo de crear una técnica de procesamiento de datos óptima y fácil de reproducir, así como también proveer información sobre los scripts que son ejecutados para completar el procesamiento de datos.

Selección de la fuente de información

La fuente que alberga los datos de interés se localiza en la página de gobierno http://www.informeseguridad.cns.gob.mx/ en dicha página web la Secretaría de Seguridad y Protección Ciudadana (SSPC), pone a disposición reportes diarios en formato pdf (Fígura 1). Los reportes están clasificados con base en la siguiente temática:

- Homicidios dolosos grupo interinstitucional
- Homicidios dolosos fuentes abiertas
- Robo de autos
- Desvió de hidrocarburos

El objeto de interés se encuentra en las columnas: "Homicidios dolosos grupo interinstitucionalz "Homicidios dolosos fuentes abiertas".

Descarga de archivos

La interfaz gráfica del micro sitio permite identificar fácilmente la fecha y el reporte que corresponde a esa fecha como lo muestra la Figura 2.

Una vez identificadas las fechas y la columna de interés, se da inicio a la descarga de los archivos pdf a través del script titulado **01_data_download_homicides_sspc_fuentesabiertas.py** o **01_data_download_homicides_sspc_gpointerinstitucional.py** según sea el caso; ambos archivos ubicados en el directorio **homicides-mx-data/data_download/**

Fecha	Homicidios dolosos Grupo interinstitucional	Homicidios dolosos Fuentes abiertas	Robo de autos	Desvío de hidrocarburos
17 de Septiembre de 2020				
16 de Septiembre de 2020				
15 de Septiembre de 2020				
14 de Septiembre de 2020				

Figura 1: Estructura principal página oficial.

Fecha	Homicidios dolosos Grupo interinstitucional	Homicidios dolosos Fuentes abiertas
17 de Septiembre de 2020		
16 de Septiembre de 2020		
15 de Septiembre de 2020		
14 de Septiembre de 2020		
13 de Septiembre de 2020		

Figura 2: Columnas de interés.

Fuentes abiertas

Una de las características principales de los datos provenientes de la columna fuentes abiertas es la desagregación por hombre, mujer y no identificado sin embargo fue hasta el mes de febrero de 2019 donde se empezó a considerar esta desagregación, por tal motivo los registros correspondientes al mes de enero de 2019 en las columnas "Hombre", "Mujerz "No Identificado"se establecieron como "NaN".

Grupo interinstitucional

Para esta sección, la página oficial de gobierno inició a reportar los homicidios desagregados por .[Es]tados.[el] 3 de abril de 2019,anterior a esa fecha los reportes se encuentran en el formato que se aprecia en la Figura 3.

Por tal motivo para los meses de enero, febrero y marzo de 2019 no existe un registro en el directorio ni en la base de datos.

Recapitulando, la figura 3 muestra el tipo de reporte que el informe de seguridad actualizó para el rango de fechas del 1 de enero de 2019 al 2 de abril de 2019, mientras que la figura 4 expone el cambio que se estableció.

Figura 3: Grupo interinstitucional enero 2019 - abril 2019

Entidades	Septiembre 15
Aguascalientes	0
Baja California	**10**
Baja California Sur	0
Campeche	1
Chiapas	4
Chihuahua	3
Ciudad de México	3
Coahuila	0
Colima	2
Durango	1
Estado de México	3
Guanajuato	**13**
Guerrero	1
Hidalgo	0
Jalisco	**6**
Michoacán	**6**

Entidades	Septiembre 15
Morelos	1
Nayarit	0
Nuevo León	0
Oaxaca	2
Puebla	2
Querétaro	0
Quintana Roo	1
San Luis Potosí	1
Sinaloa	2
Sonora	2
Tabasco	0
Tamaulipas	3
Tlaxcala	2
Veracruz	**7**
Yucatán	0
Zacatecas	0
Total	**76**

Figura 4: Grupo interinstitucional abril 2019 - diciembre 2019

Diagramas UML

Con el fin de abstraer y exponer el ciclo del sistema de extracción de datos y procesamiento, las siguientes secciones muestran diagramas a través del lenguaje unificado de modelado (UML), presentando diagramas de actividades para ilustrar la naturaleza dinámica del sistema mediante el modelado de flujo de las actividades. Típicamente los diagramas de actividad en la Ingeniería de Software son utilizados para modelar el flujo de trabajo interno de una operación.
En rasgos generales la Figura 5 corresponde al diagrama de actividad del sistema completo desde la descarga de archivos hasta el almacenamiento de la base de datos limpia y lista para su posterior uso.

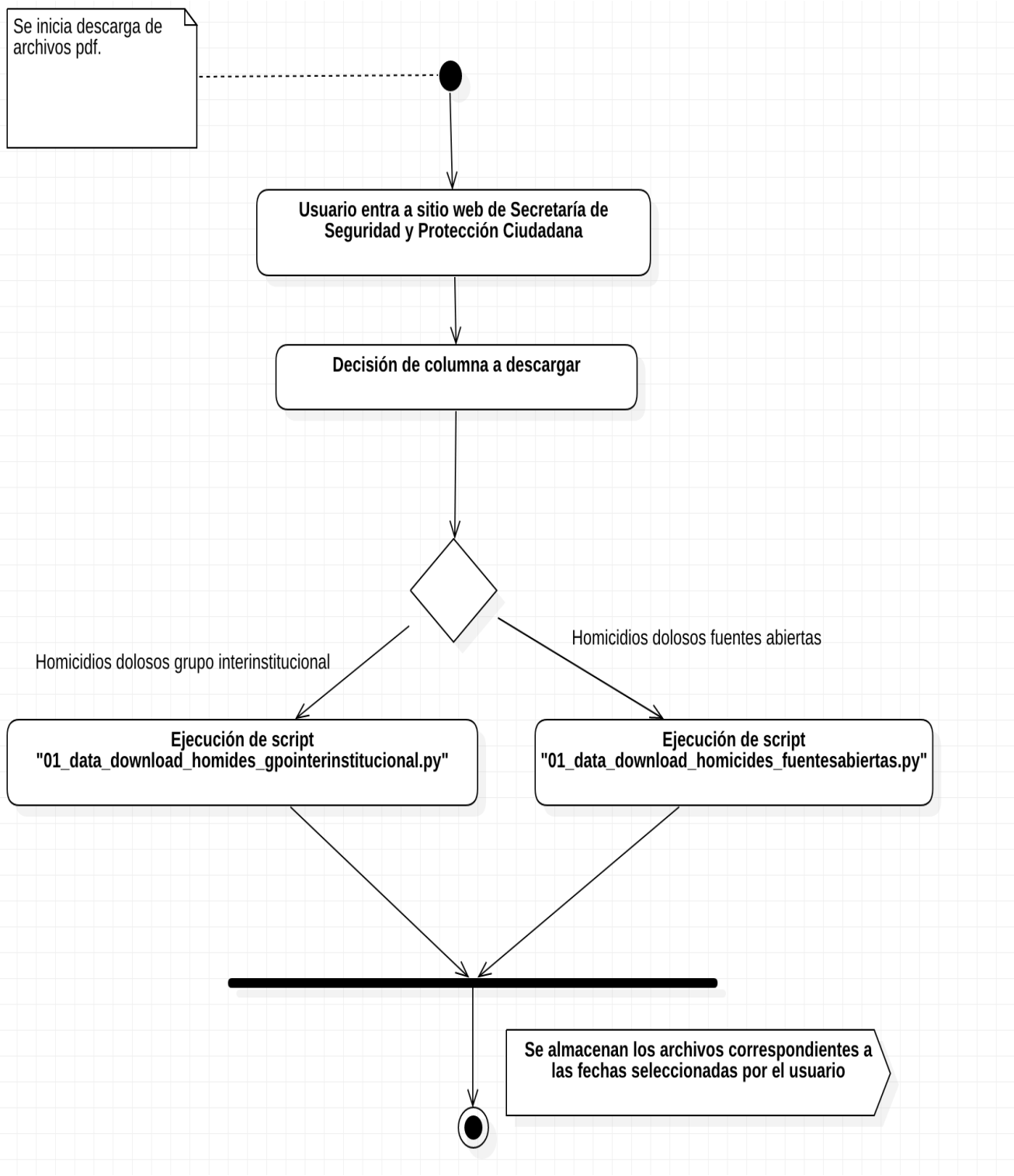


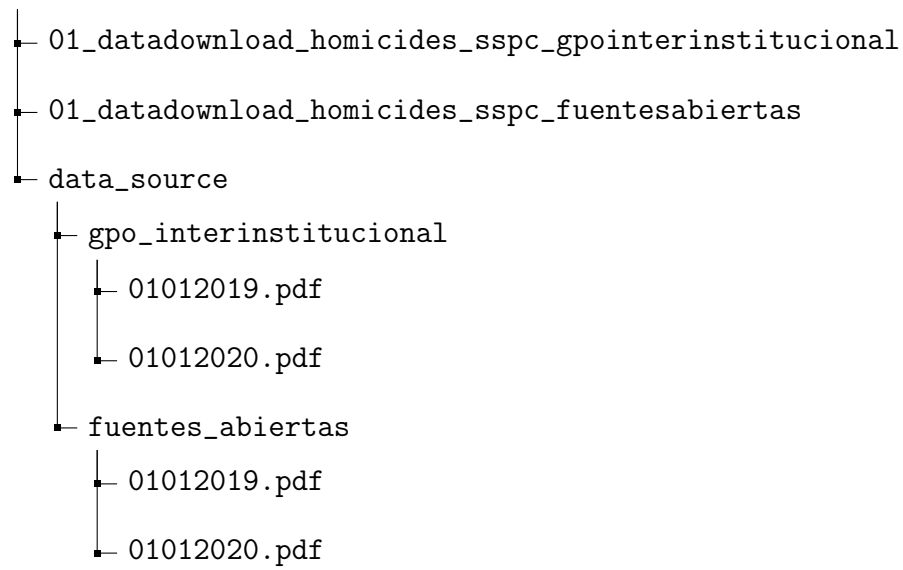
Figura 5: Diagrama de Actividad-Sistema de datos

Los códigos están programados de una manera intuitiva para que el usuario le ataña únicamente ejecutar los scripts que se indican. Si es de interés conocer a fondo el funcionamiento de cada función, la documentación relacionada a los scripts se encuentra en la dirección (Repositorio homicidios) .

Almacenamiento

Retomando el proceso de descarga,una vez que se ha ejecutado el script
01_data_download_homicides_gppinterinstitucional.py y/o
01_data_download_homicides_fuentes_abiertas.py los archivos extensión .pdf se almacenaran en en subdirectorios de la carpeta homicides-mx-data/data_download/data_source/

El diagrama siguiente, muestra una estructura de la organización del directorio homicides-mx-data/data_download/.



Extracción de datos

La extracción de los datos provenientes de los archivos pdf se realiza con el script
02_data_extraccion_homicides_sspc fuentesabiertas.py o
02_data_extraction_homicides_sspc gpointerinstitucional según sea el caso.

La tarea principal de estos scripts es extraer los datos y la estructura de las tablas de la forma más precisa posible para que la labor de limpieza se pueda realizar a través de un proceso estandarizado.
Los detalles sobre el funcionamiento de cada script se pueden consultar a través de la documentación en el repositorio.

Cuando culmina el proceso de extracción, se generan archivos equivalentes a los pdf pero ahora en formato .csv y con una estructura tabular, es decir, el archivo csv generado es una replica de la tablas que se albergan en los archivos pdf; estos archivos de extensión csv se almacenan en el directorio homicides-mx-data/data_raw/

Con el objetivo de tener un mayor control sobre cada proceso es menester aclarar que el directorio que hospeda los archivos transformados a formato csv se clasifican en subdirectorios que

pueden ser fácilmente identificados a través del nombre. La siguiente figura expone la clasificación y organización del directorio homicides-mx-data/data_raw.

Como se puede apreciar en la estructura, el directorio contiene dos subdirectorios principales que corresponden a la categoría fuentes abiertas y grupo interinstitucional, a su vez cada uno de estos subdirectorios contiene otras carpetas que se en listan a continuación.

- 2019 : Contiene desglosado por meses los registros diarios del año 2019.
- 2020 : Contiene desglosado por meses los registros diarios del año 2020.
- enero,febrero ... : Contiene los archivos por día correspondientes a cada mes y cada año.
- monthly : Alberga la concatenación de todos los días del año descargados.
- 2019_2020 : Data frame, resultado de concatenar todos los registros del año 2019 y 2020 registrados.

Los archivos por mes que se encuentran en el directorio monthly de cada año, se identifican con un índice que corresponde al mes, para ser fácil de identificar no solamente por el nombre sino también por el index.

```
data raw
├─ fuentes abiertas
│  ├─ 03_data_concatenation_homicides_sspc_fuentes_abiertas
│  ├─ 2019
│  │  ├─ january
│  │  │  ├─ 01_01_2019.csv
│  │  │  └─ ...
│  │  ├─ february
│  │  ├─ ...
│  │  ├─ december
│  │  └─ monthly
│  │     ├─ 01_january_2019
│  │     ├─ 02_february_2019
│  │     ├─ ...
│  │     ├─ 12_december_2019
│  │     └─ df_homicides_daily_2019_sspc_fuentesabiertas.csv
│  ├─ 2020
│  │  ├─ january
│  │  │  ├─ 01_01_2020.csv
│  │  │  └─ ...
│  │  ├─ february
│  │  ├─ ...
│  │  ├─ december
│  │  └─ monthly
│  │     ├─ 01_january_2020
│  │     ├─ 02_february_2020
│  │     ├─ ...
│  │     ├─ 12_december_2020
│  │     └─ df_homicides daily_2020_sspc_fuentesabiertas.csv
│  └─ 2019_2020
│     └─ df_homicides_daily_2019_2020_sspc_fuentesabiertas.csv
└─ gpo_interinstitucional
```

```
data_raw
└── gpo_interinstitucional
    ├── 2019
    │   ├── april
    │   ├── may
    │   ├── ...
    │   ├── december
    │   └── monthly
    │       ├── 04_april_2019
    │       ├── 02_february_2019
    │       ├── ...
    │       ├── 12_december_2019
    │       └── df_homicides_daily_2019_sspc_ gpointerinstitucional.csv
    ├── 2020
    │   ├── january
    │   ├── february
    │   ├── ...
    │   ├── december
    │   └── monthly
    │       ├── 01_january_2020
    │       ├── 02_february_2020
    │       ├── ...
    │       ├── 12_december_2020
    │       └── df_homicides_daily_2020_sspc_ gpointerinstitucional.csv
    └── 2019_2020
        └── df_homicides_daily_2019_2020_sspc_ gpointerinstitucional.csv
```

Limpieza de datos

La parte modular del sistema de extracción y limpieza de datos reside en esta sección, para cada columna se siguió una serie de pasos diferentes que se pueden apreciar de mejor manera en los diagramas de actividades de las figuras 6 y 7.

Fuentes abiertas

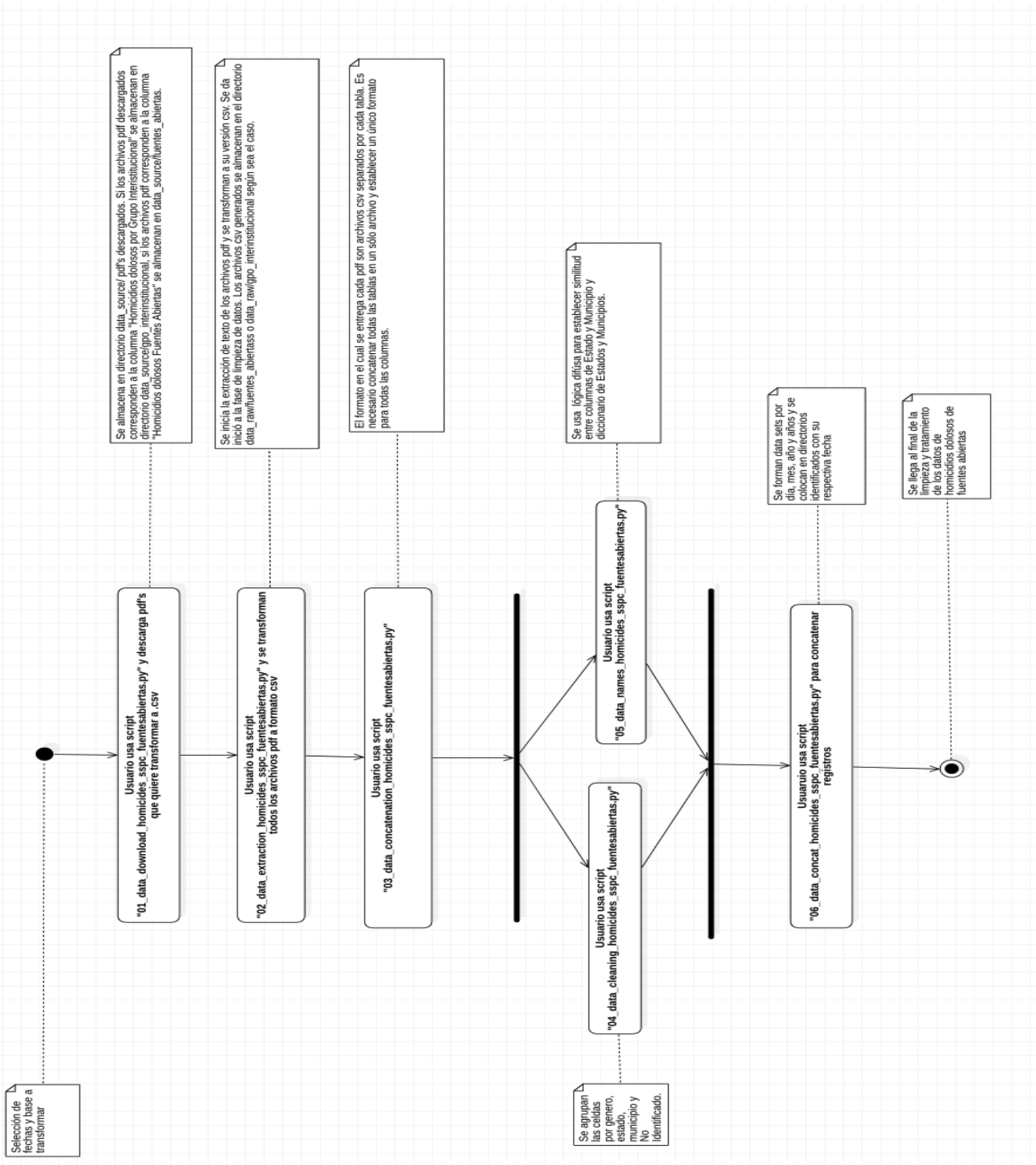


Figura 6: Diagrama de actividades limpieza fuentes abiertas

Grupo interinstitucional

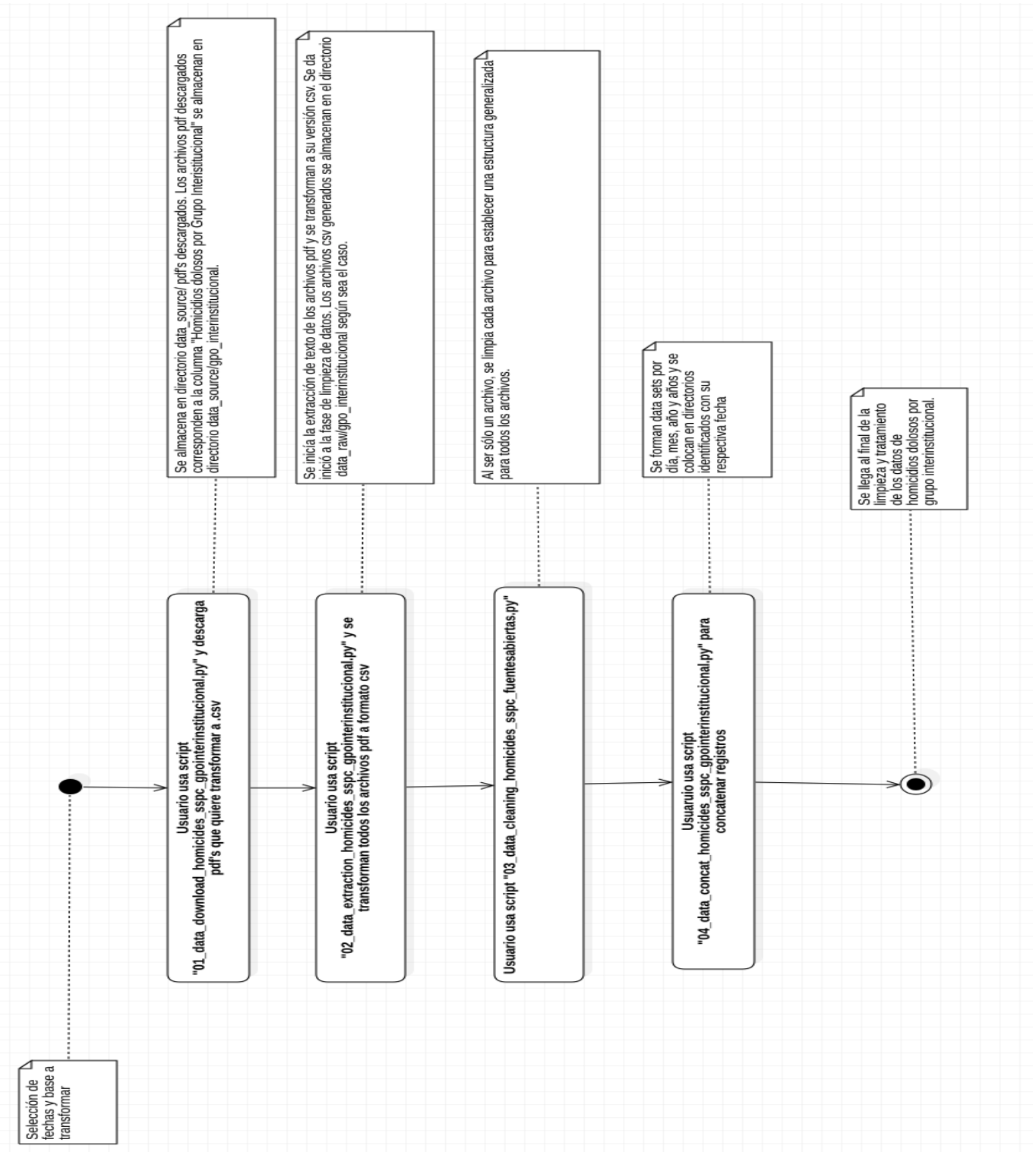


Figura 7: Diagrama de actividades limpieza grupo interinstitucional

Aseguramiento de la calidad

Con el fin de mantener y asegurar la calidad a través de la consistencia, integridad, cohesión y coherencia de las bases de datos; cada etapa mencionada anteriormente fue sometida a revisiones meticulosas por diferentes miembros del equipo si se llegaba a identificar alguna anomalía o error en alguna etapa del sistema se procedió a registrar cada punto atípico en la bitácora de datos ubicada en el directorio data_validation.

Contacto

Asegurar un proceso óptimo para generar un sistema de extracción y limpieza de datos conlleva el deber de asegurar en cada paso código suficientemente entendible para que proceso pueda ser reproducible y eventualmente automatizado. Si el usuario de este manual desea aportar información de valor para la mejora o bien ayudar a identificar errores, es posible contactar a los y las desarrolladoras del sistema a través de la página web (PADeCI).