Reporte Técnico Extracción de Datos

Secretaría de Seguridad y Protección Ciudadana (SSPC)

Yadira Elizabeth Peralta Torres
Fernando Alarid Escudero
Mariana Consuelo Fernández Espinosa
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.

Selección de la fuente de información

La fuente que alberga los datos de interés se localiza en la página de gobierno http://www.informeseguridad.cn en dicha página web la Secretaría de Seguridad y Protección Ciudadana (SSPC), pone a disposición reportes diarios en formato pdf. 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 interinstitucional"y "Homicidios dolosos fuentes abiertas".

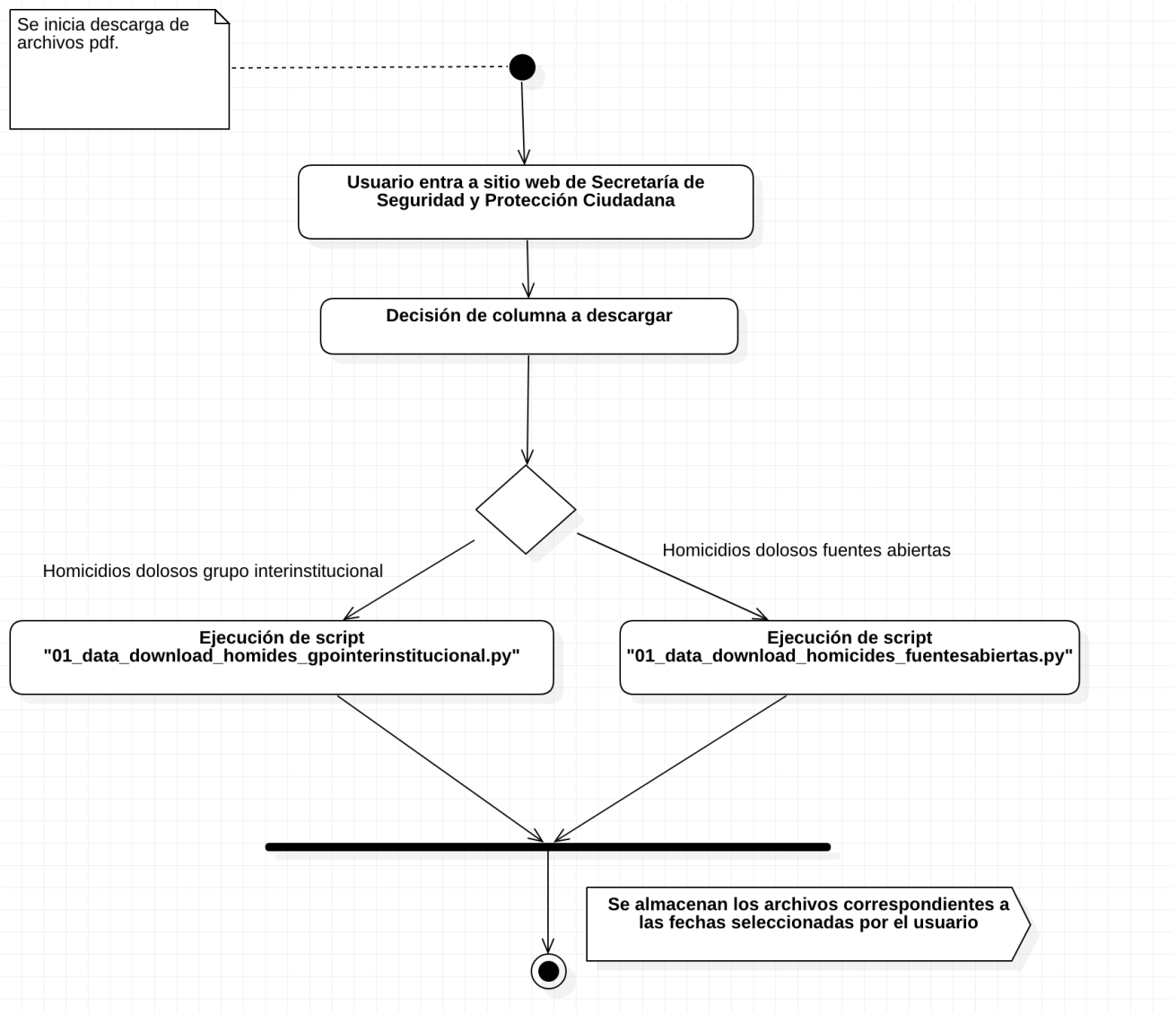
Fecha	Homicidios dolosos Grupo interinstitucional	Homicidios dolosos Fuentes abiertas	Robo de autos	Desvío de hidrocarburos
17 de Septiembre de 2020				
16 de Septiembre de 2020				
15 de Septiembre de 2020				
14 de Septiembre de 2020				

Descarga de archivos

La interfaz gráfica del micro sitio permite identificar fácilmente la fecha y el reporte que corresponde a esa fecha

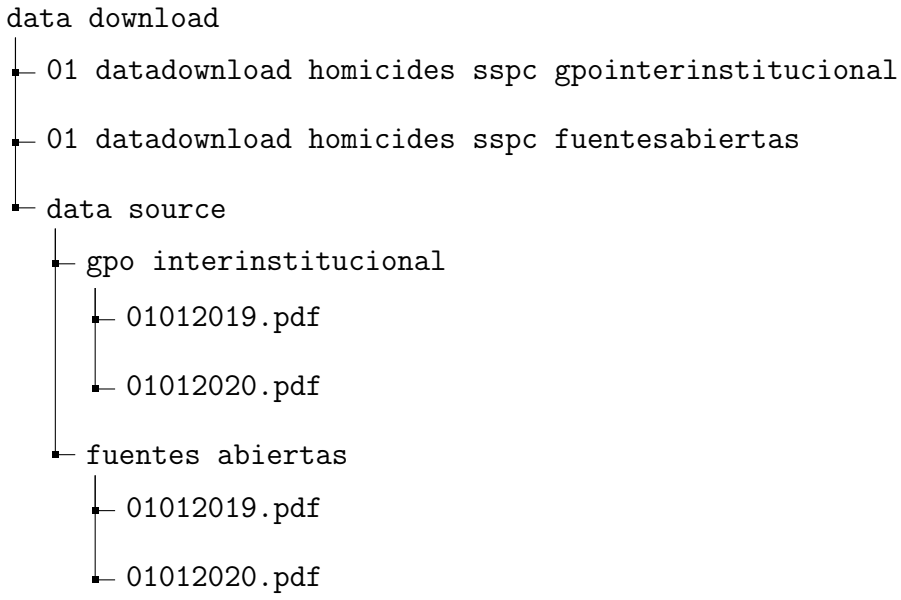
Fecha	Homicidios dolosos Grupo interinstitucional	Homicidios dolosos Fuentes abiertas
17 de Septiembre de 2020		
16 de Septiembre de 2020		
15 de Septiembre de 2020		
14 de Septiembre de 2020		
13 de Septiembre de 2020		

Cuando se han identificado las fechas y la columna de interés , los scripts ubicados en el directorio data download servirán para iniciar la descarga de archivos.



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 (link).

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 data download.



Extracción de datos

En esta sección, 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 este script es extraer los datos y la estructura de las tablas de la forma más precisa posible para que la tarea de limpieza se pueda realizar a través de un proceso estandarizado.
Los detalles sobre el funcionamiento de cada script se pueden consultar a traves de la documentación del 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 data raw/ y en el subdirectorio correspondiente.

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 data raw.

```
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
```

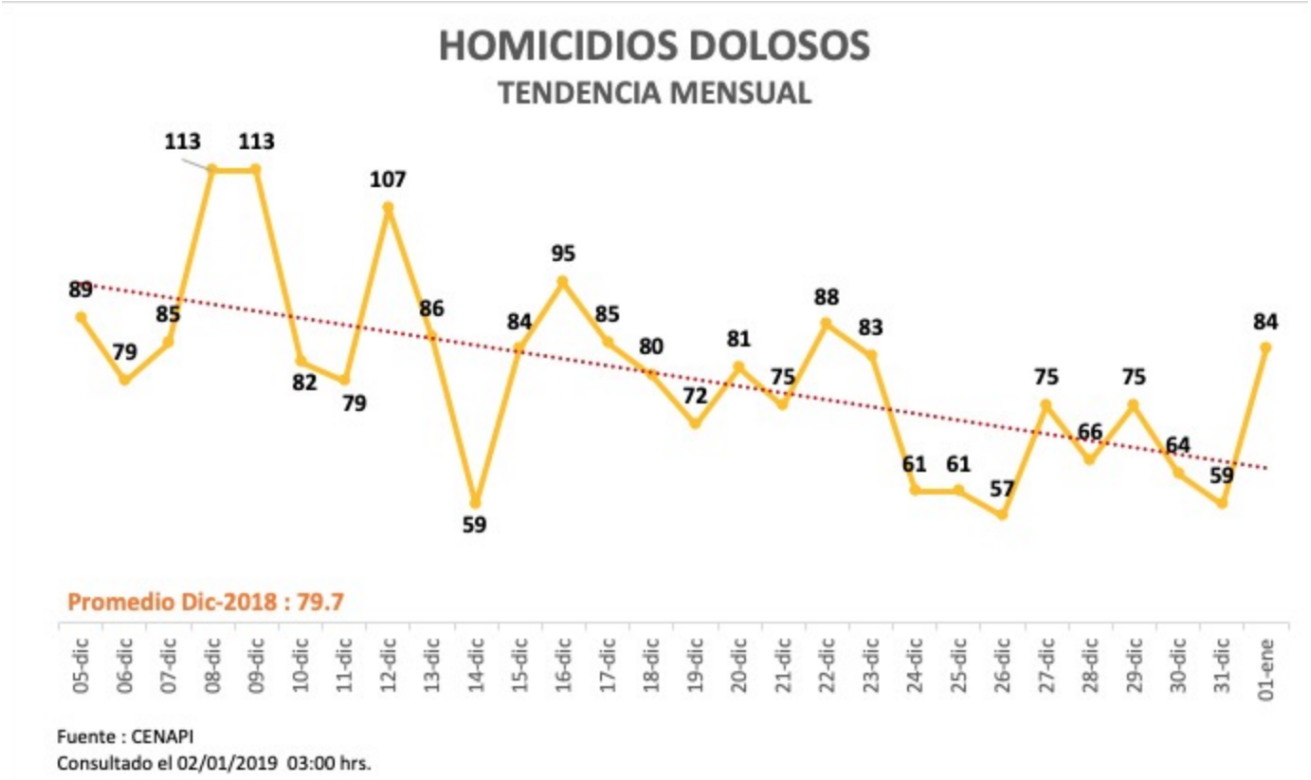
Fuentes abiertas

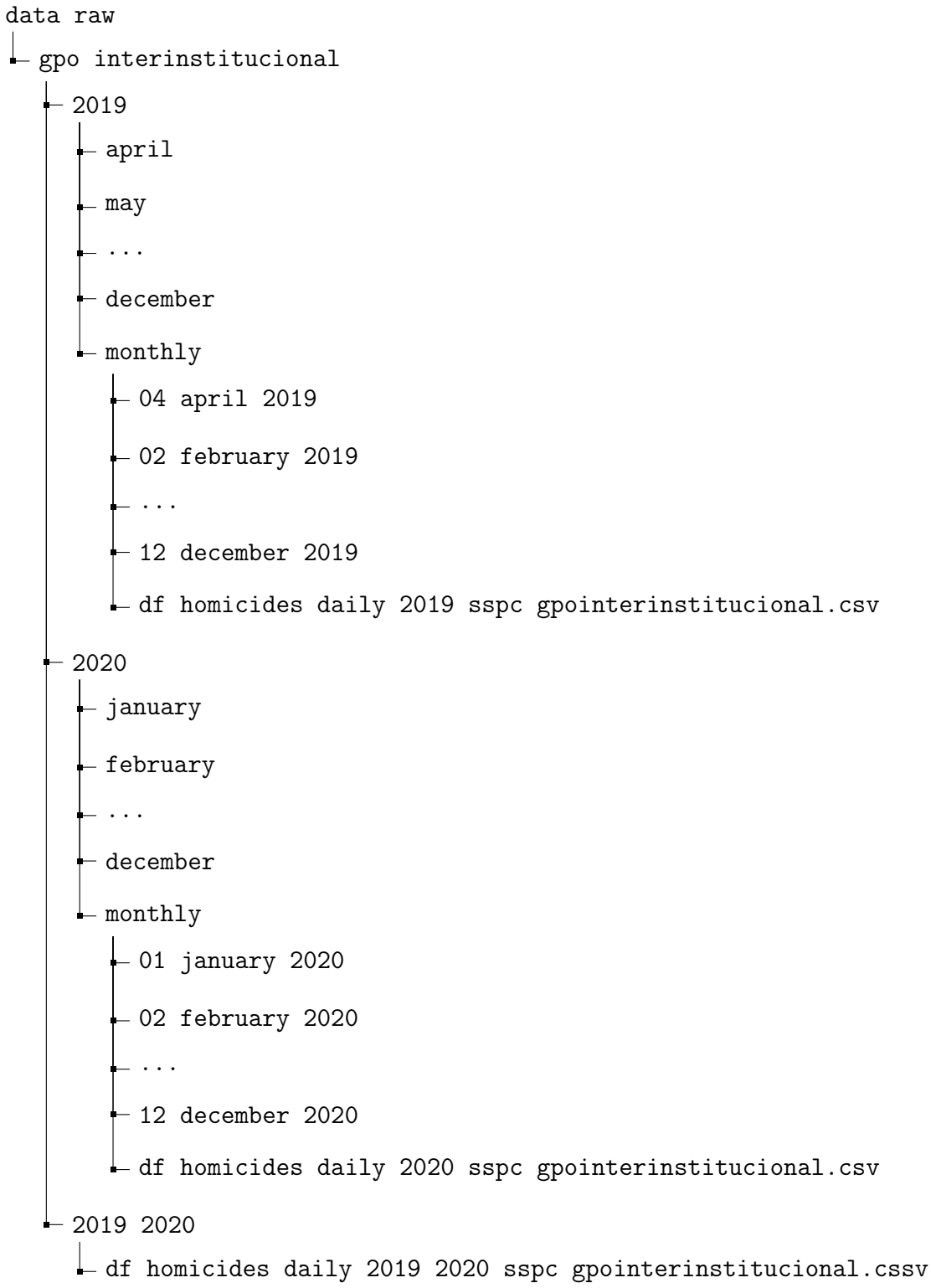
Una de las características principales de los datos provenientes de la columna fuentes abiertas es la desagregación por hombre, mujer y no identificado sin embargo fue hasta el mes de febrero de 2019 donde se empezó a considerar esta desagregación por tal motivo los registros correspondientes al mes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

Grupo interinstitucional

El subdirectorio grupo interinstitucional sostiene una estructura de ficheros similar a fuentes abiertas. En el siguiente diagrama se aprecia de mejor manera la organización.

Para esta sección el micrositio inició el reporte desagregado por Estados el 3 de abril de 2019, por tal motivo para los meses de enerero, febrero y marzo de 2019 no existe un registro en el directorio.
La siguiente imagen muestra el tipp 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.





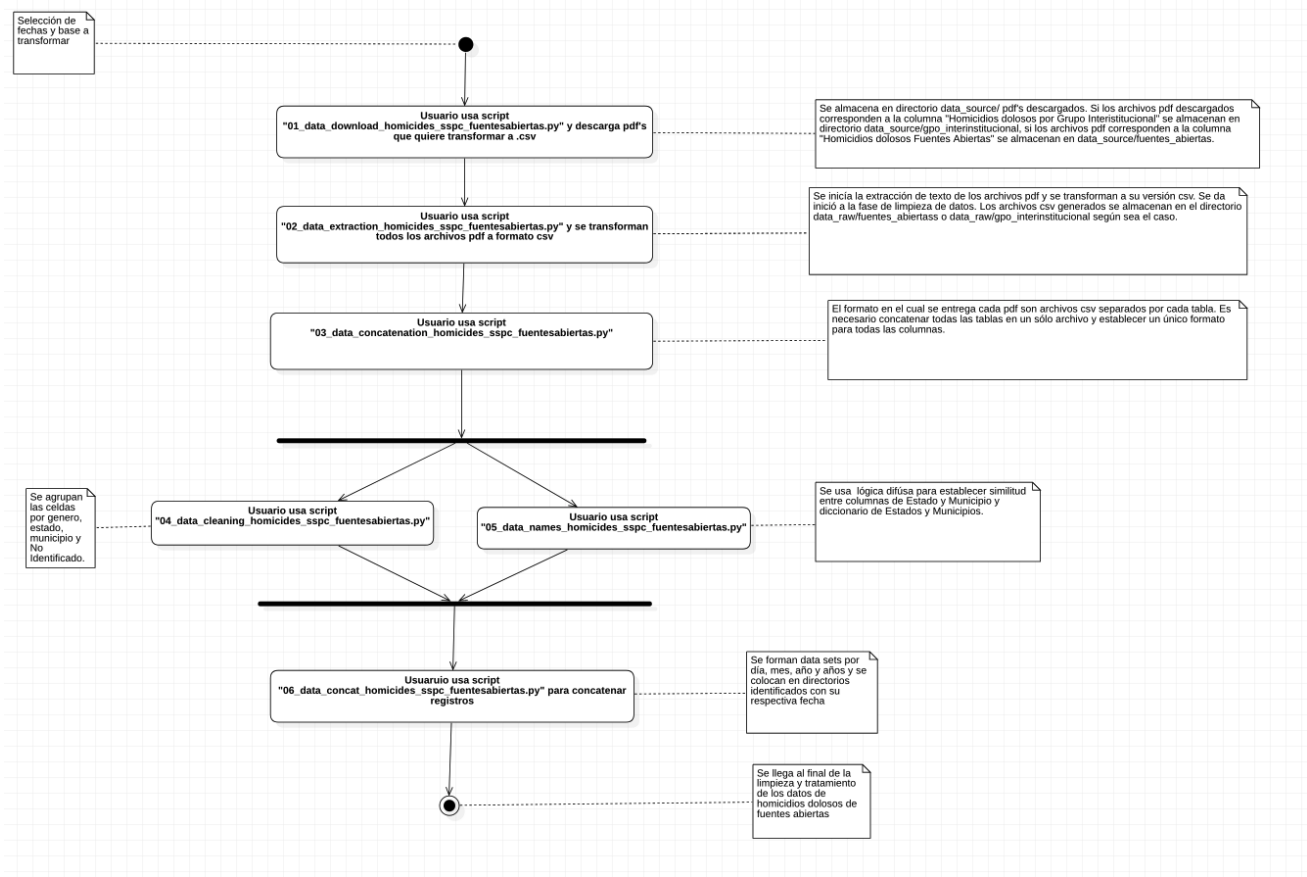
2019 2020

El directorio alberga como lo indica el nombre, el data set correspondiente al año 2019 y 2020 por registro diario.

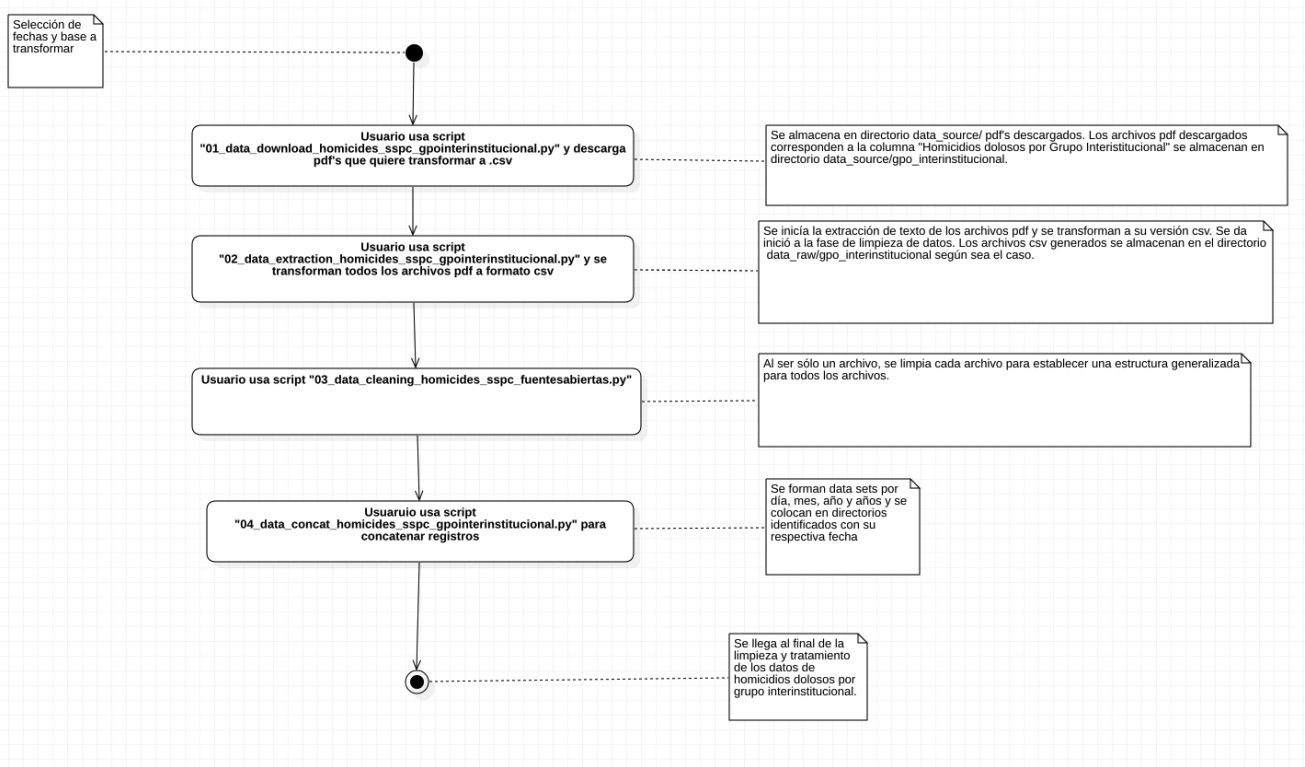
Limpieza de datos

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.

Fuentes abiertas



Grupo interinstitucional



Aseguramiento de la calidad

Conforme a la norma ISO 9000: 2000, la calidad se podría definir como “el grado en el que un conjunto de características inherentes cumple con los requisitos, esto es, con la necesidad o expectativa establecida, generalmente implícita u obligatoria".

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 anómalia o error en alguna etapa del sistema se procedio a registrar cada punto atípico en la bitacora de datos ubicada en el directorio data validation.

Conclusión