Equipo de Estado Abierto

Ciudad de México, 9 de mayo de 2025

Nota Metodológica 2
Limpieza y testado de las Solicitudes de Información Pública (SIP) 2024

Este documento describe el procedimiento implementado para la limpieza y el testado de bases de datos trimestrales relacionadas con Solicitudes de Información Pública (SIP) correspondientes a 2024. El objetivo principal fue garantizar la calidad de los datos para su análisis y reporte, procurando también la protección de información personal de las personas solicitantes. El enfoque incluyó procesos específicos de filtrado, depuración, eliminación de información no relevante y anonimización de datos como números telefónicos, nombres y otra información que identifique o haga identificable a las personas.

Etapa 1: Limpieza de las bases de datos

1. Filtrado de solicitudes relevantes

El primer paso en el proceso de limpieza consistió en seleccionar únicamente las SIP y eliminar aquellas que correspondían a solicitudes de acceso a datos personales. Este filtrado se realizó para garantizar que el análisis se enfoque exclusivamente en las solicitudes de información pública.

2. Eliminación de duplicados

Para evitar duplicidades de las solicitudes, se revisaron los registros de la base de datos y se eliminaron aquellos que correspondían a solicitudes idénticas. Este proceso consideró: la coincidencia exacta en variables clave como el número de folio, la fecha de presentación y el contenido textual de la solicitud.

3. Eliminación de variables no relevantes

Una vez que la base de datos fue depurada, se identificaron y eliminaron aquellas variables que no contribuyen directamente al análisis. Este paso permitió reducir la complejidad del conjunto de datos y centrarse en las variables de mayor interés, como el texto de la solicitud, la fecha, el folio y las clasificaciones.
Criterios para determinar relevancia:

- Variables que no aportan información sustancial para los objetivos del estudio (e.g., identificadores administrativos internos.
- Variables que podían derivarse de otras ya presentes en el conjunto de datos.

Etapa 2: Testado de información personal

El testado tuvo como propósito anonimizar cualquier información personal presente en las solicitudes, con el fin de proteger la privacidad de las personas. Este procedimiento incluyó la identificación, sustitución y validación de información de carácter personal, por la leyenda “T E S T A D O”. La anonimización se logró a través de la identificación sistemática de datos personales mediante patrones predefinidos y su posterior sustitución por texto genérico que indica su anonimización.

Datos Personales Testados

Durante el proceso, se definieron y aplicaron patrones específicos para identificar y testear los siguientes tipos de datos personales, los cuales tienen una alta probabilidad de exposición:

1. Nombres completos:

Se diseñaron patrones para reconocer nombres típicos en español, incluyendo combinaciones comunes de nombre y uno o dos apellidos.
Estrategia de detección: Se identificaron palabras que comienzan con mayúsculas seguidas de minúsculas, considerando estructuras gramaticales frecuentes de nombres propios.

2. Correos electrónicos:

Se buscaron direcciones de correo con formato estándar (e.g., usuario@dominio.com).
Estrategia de detección: Se incluyeron tanto dominios genéricos (e.g., gmail.com) como específicos. (e.g., institucionales).

3. Números telefónicos:

Se detectaron números en formatos locales e internacionales, incluyendo secuencias de 10 dígitos y aquellas con prefijos (+52, etc.).
Estrategia de detección: Se consideraron variaciones con espacios, guiones y otros separadores comunes.

4. CURP y RFC:

Se buscaron estas claves oficiales mediante patrones alfanuméricos específicos, únicos en su estructura y longitud.
Estrategia de detección: Los patrones incluyeron validaciones de formato estándar para tomar en cuenta la precisión.

5. Códigos postales:

Secuencias numéricas de cinco dígitos asociadas a ubicaciones geográficas específicas.

Estrategia de detección: Se consideraron cuando aparecían en combinación con otros datos como direcciones.

6. Direcciones completas:

Se analizaron frases que incluyeran palabras clave como “Calle”, “Avenida”, “Colonia” o referencias numéricas, seguidas de códigos postales.
Estrategia de detección: Los patrones fueron diseñados para identificar direcciones comunes y estructuradas.

Proceso de testado

1. Definición de patrones:

Se diseñaron patrones basados en expresiones regulares (regex), capaces de identificar variaciones comunes de escritura en los datos personales, como:
Espacios y guiones en números telefónicos.
Diferencias en el uso de mayúsculas/minúsculas en nombres.
Formatos específicos de correos electrónicos, CURP y RFC.
Cada patrón fue adaptado para minimizar falsos positivos y maximizar la precisión.

2. Aplicación de patrones a las columnas relevantes:

Se identificaron las columnas con mayor probabilidad de contener datos personales, como “SOLICITUD” y “RESPUESTA”, de manera que, se aplicó el proceso de testado a las anteriores.
A través de un proceso automatizado, cada patrón fue aplicado iterativamente para detectar y anonimizar los datos personales.

3. Sustitución de datos personales:

Todos los elementos detectados se reemplazaron con el texto genérico “T E S T A D O”.
Es importante mencionar que la sustitución homogénea evita inconsistencias que podrían afectar el análisis estadístico.

4. Verificación de resultados:

Se implementó un control de calidad mediante la revisión de muestras aleatorias de los datos testados.
Se compararon registros antes y después del proceso para asegurar que no quedaran datos personales sin anonimizar.
Además, se revisó que los patrones no eliminaran información relevante.