

# Limpieza de Datos

Alvaro Machicado   Diego Santos   Maximiliano Cabezón Alvarez

*Calidad de Datos - Departamento de Computación - Universidad de Buenos Aires*

## 1. Análisis del dataset

En el análisis preliminar del dataset `GSAF5.xml` sobre las columnas `Country` y `Type` encontramos las siguientes inconsistencias:

- Sobre `Country`:

- Países duplicados por espacios en blanco
- Países duplicados por errores de ortografía.
- Múltiples formas de referirse al mismo país.
- Referencias geográficas en lugar del nombre del país. Ej: Gulf Bay
- Rango de países.
- Continentes en lugar de países.
- Océanos en lugar de países.

.

- Sobre `Type`:

- Tipo que no pertenecen a la lista de tipos posibles.
- Descriptivos del tipo de incidente en lugar del Tipo de incidente.

## 2. Limpieza del dataset

Para limpiar el dataset se utilizó la herramienta OpenRefine. Las acciones que fuimos tomando fueron:

### 2.1. *Para Type*

Nos apegamos al código de colores ya que nos pareció consistente:

- Eliminamos 59.000 filas que tenían todos sus campos en blanco.
- Corregimos 5 casos en los que pudimos aplicar la regla de colores
- UNCONFIRMED, UNDER INVESTIGATION, UNVERIFIED los etiquetamos como QUESTIONABLE
- La mayoría de los casos que se marcaban como INVALID los pusimos bajo QUESTIONABLE pero en casos particulares los marcamos como UNPROVOKED porque estaban asociados al color naranja.

### 2.2. *Country*

Utilizamos los siguientes criterios:

- Usamos **Cluster** para unificar países similares
- Renombramos columna **Country** a **Site**: para evitar perder información borrando aquellos casos en los cuales el atributo no era un país.
- Agregamos la columna **Site Type**: esta columna se encarga de identificar si lo que tenemos en la columna **Site** es un país u otra cosa (océanos, entre países, etc). Para eso, los valores que puede tomar son **Country** y **Other**.
- Completamos manualmente 50 casos que tienen el campo **Site** en blanco: la mayoría terminaron en estado UNKNOWN por no tener información suficiente.

Aplicando todas esas correcciones pudimos extraer un dataset consistente con el cual trabajar.

## 3. Dataset obtenido

Dentro de la carpeta **/resultados** adjuntamos los siguientes archivos:

**GSAF5.xlsx** contiene la base de datos obtenida luego de aplicar la limpieza de los datos.

En ella agregamos tablas pivot donde contabilizamos los resultados en las pestañas:

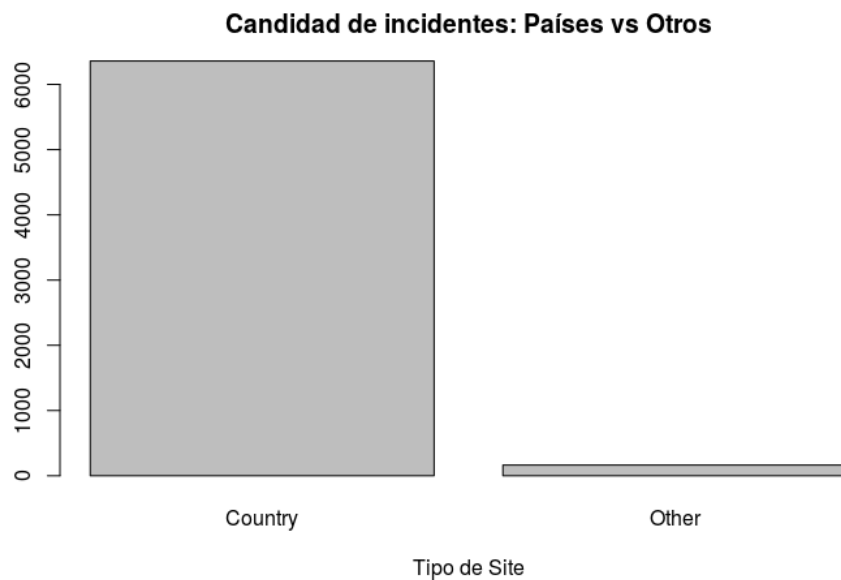
- Cantida de Incidentes por Tipo de incidente

- Cantidad de incidentes por País
- Cantidad de incidentes agrupados por País por Tipo
- Cantidad de incidentes agrupados por Tipo por País

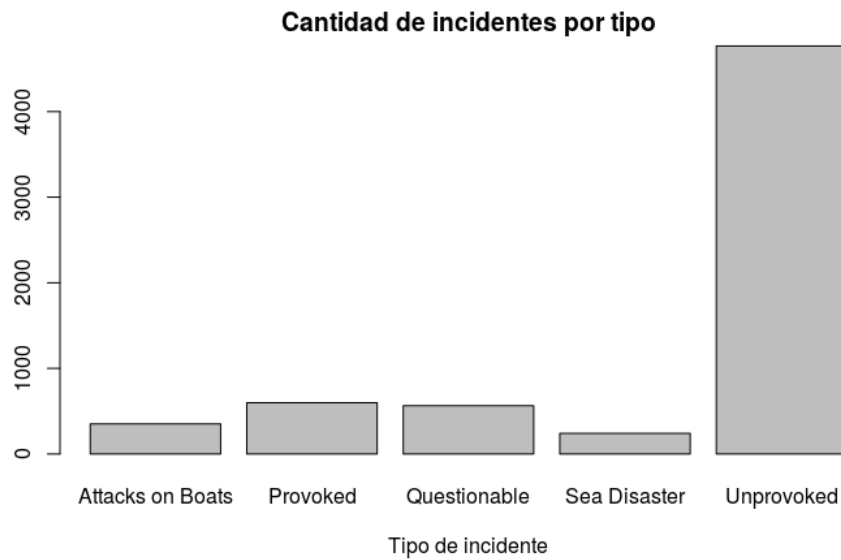
También agregamos en la carpeta resultados `GSAF5-histoy.json`, el archivo JSon autogenerado por Open Refine al aplicar la limpieza.

## 4. Gráficos

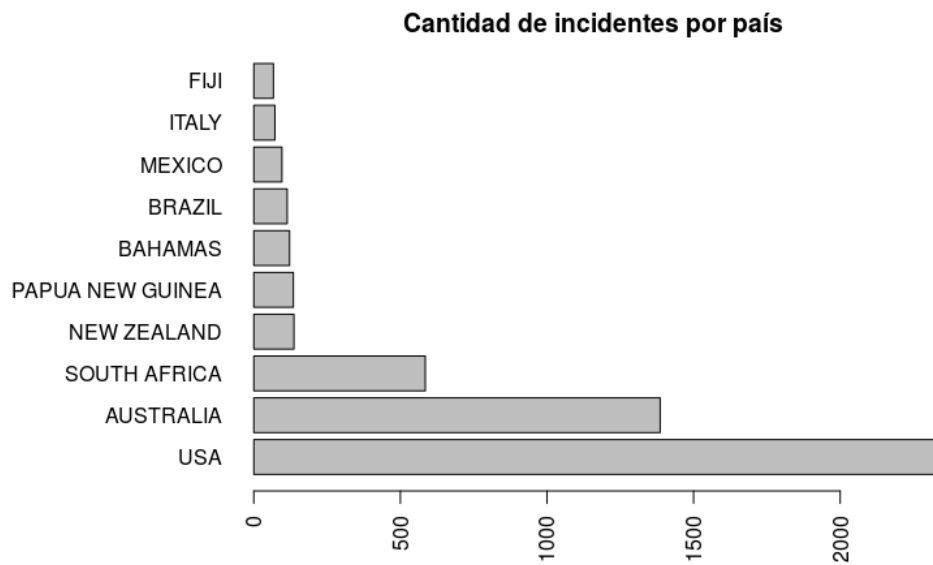
Con la clasificación adicionada sobre el lugar del accidente obtuvimos la distribución entre países y territorios aproximados:



De la limpieza de los atributos de la columna Type obtuvimos la siguiente distribución de causas.



Los 10 países con más incidentes de tiburones son:



## 5. Adicionales

Agregamos en la carpeta **scripts** los scripts que usamos para generar los gráficos de este informe.