MÓDULO 3

LAS 5PS DE LA AGENDA 2030 Y SU ARTICULACIÓN AL DISCURSO PERIODÍSTICO

PERIODISMO DE DATOS
PARA EL DESARROLLO SOSTENIBLE

EN EL MUNDO DE LOS DATOS:
LIMPIAR

Algunas bases de datos, ya sea por la forma en que fueron capturados los registros o, incluso, por un problema de codificación entre sistemas, pueden contener valores que afectan la manera en que se puede trabajar de forma eficiente la información. La limpieza de datos busca que esa misma información sea presentada de forma tal que **su procesamiento no represente un problema por una falla en la interpretación de los valores o por duplicación de registros.**

Día a día crece la cultura de entregar bases de datos "limpias"; sin embargo, no deja de ser necesario conocer los aspectos fundamentales de la Limpieza de Datos. En esencia todas las bases de datos que son publicadas por un organismo público o tienen algún reconocimiento oficial, deberían mostrar su información en perfectas condiciones, pero no siempre sucede de esa forma, de tal suerte que se requiere aplicar algún tipo de limpieza a la base para que esté en condiciones de ser usada.

ERRORES COMUNES EN LAS BASES DE DATOS

Hace tiempo Christopher Groskopf miembro del equipo de periodismo de datos de Quartz, escribió una guía muy completa sobre los distintos tipos de errores que las bases de datos podrían tener y qué hacer ante esas circunstancias. En este ejercicio Christopher agrupó cada una de ellas entorno a cuatro categorías principales:

1. Problemas que debería resolver la fuente
2. Cuestiones que debería resolver el usuario
3. Problemas que un tercero experto podría ayudar a solucionar
4. Problemas que un programador podría resolver

En cada una de estas categorías se describen algunas de las pesadillas a las cuales cualquier persona que usa datos puede llegar a enfrentarse. A continuación se rescatan algunas de ellas para ilustrar los errores comunes en las bases de datos que nos pueden llevar a la mejor manera de resolverlo:

VALORES FALTANTES

Cuidado con los valores en blanco o " null " en cualquier dataset , a menos que haya seguridad sobre lo que significan. Si los datos son anuales, ¿el dato para ese año no fue levantado? ¿Si es una encuesta, algún encuestado se rehusó a contestar la pregunta?
En cualquier momento del trabajo con datos faltantes debería preguntarse: "¿conozco el significado de la ausencia de este valor?" Si la respuesta es negativa, será necesario preguntarle a la fuente.

DATOS FALTANTES REEMPLAZADOS CON CEROS

- Peor que un dato faltante es el uso de un valor arbitrario en su lugar. Esto puede ser el resultado de un humano que no esté pensando en las consecuencias de ese uso o puede suceder como resultado de un proceso automatizado que simplemente no sabe cómo manejar valores nulos. En cualquier caso, de haber ceros en una serie de números debería preguntarse si esos valores corresponden realmente al número 0 o más bien, al significado "nada". (-1 también se usa a veces así). Si no hay seguridad, pregunte a la fuente.

- La misma precaución debería valer para otros valores no-numéricos donde un 0 pueda ser representado de otra manera. Por ejemplo, un falso 0 para una fecha suele ser representado como 1970-01-01T00:00:00 Z o 1969-1231T24:59:59 Z, que es el **comienzo del registro de tiempo en Unix** . Un falso 0 para una ubicación puede ser representado como 0°00'00.0"N+0°00'0 0.0"E o simplemente 0°N 0°E, que es un punto en el Océano Atlántico justo al sur de Ghana, frecuentemente llamado **Null Island.**

FILAS O VALORES QUE ESTÁN DUPLICADOS

Si la misma fila aparece en el dataset más de una vez, debería averiguar por qué. A veces no necesita ser una fila entera. Algunos datos de financiamiento de campañas incluyen "correcciones" que usan los mismos identificadores únicos que la transacción original. Si no sabía eso, entonces cualquier cálculo hecho con los datos sería incorrecto. Si hay algo que parezca debe ser único, mejor verificar que lo sea. Si descubre que no lo es, mejor preguntar a la fuente.

LA ORTOGRAFÍA ES INCONSISTENTE

- No sólo hay que fijarse en los nombres de la gente, ese es uno de los sitios donde más difícil es hallar errores. En lugar de esto, busque lugares donde los nombres de estados o ciudades no sean consistentes. (Los Angelos es un error muy común).

- Si encuentra errores de este tipo, puede estar seguro de que los datos fueron compilados o editados a mano y esa es razón suficiente para guardar cierto escepticismo. Los datos editados a mano son los más proclives a fallas. Esto no significa que no debería usarlos pero habrá que corregirlos manualmente o publicarlos como errores en el reporte.

- La herramienta de **Open Refine** para **agrupar texto** puede ayudar en ese proceso sencillo y eficiente al sugerir coincidencias cercanas entre valores inconsistentes en una columna (por ejemplo, igualando Los Angelos con Los Ángeles). Es importante **documentar los cambios**, para garantizar un **buen origen de los datos.**

Si está en México, Colombia o Costa Rica, pase por el ecosistema de datos de **www.datarepublica.org** y explore un poco el panorama de la disponibilidad de datos para cada Objetivo de Desarrollo Sostenible (ODS) en estos países.

DATAREPÚBLICA

Cepei
Knowledge factory for the world we deserve