Métricas de calidad para validar los conjuntos de datos abiertos públicos gubernamentales

Tesis Doctorales

AUTORA:

ROXANA MARTÍNEZ
maria.martinez@info.unlp.edu.ar
ing.roxana.martinez@gmail.com

Facultad de INFORMÁTICA
UNIVERSIDAD NACIONAL DE LA PLATA

UNIVERSIDAD NACIONAL DE LA PLATA

DIRECCIÓN:
Directora: Rocío Rodríguez
Co-Directora: Claudia Pons
Asesor Científico: Pablo Vera

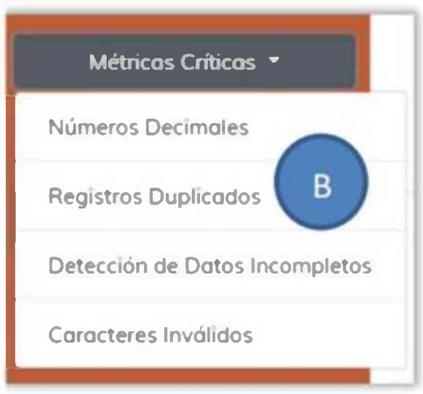
Título Obtenido: "Doctora en Ciencias Informáticas"

Fecha de exposición: 29 de marzo de 2022
Publicación: 2022

Jurado de Tesis Doctoral:
Dr. Mario Piattini (España)
Dra. Elsa Estévez (Argentina)
Dr. Pablo Fillotrani (Argentina);

PALABRAS CLAVE:

Datos Abiertos, Datos Públicos, Gobierno Abierto, Métricas de Calidad de Datos.



MOTIVACIÓN:

- Los datasets brindados en los portales de datos abiertos no cumplen con un estándar en su contenido.
- Si bien existen principios y criterios internacionales de apertura de datos, no hay un enfoque en el análisis del contenido de éstos.
- Existen problemas que pueden ser mitigados con anterioridad a su publicación, en aspectos estructurales y de formatos (Interoperabilidad).
- Los conjuntos de datos no siempre son suficientes o bien legibles.
- Es necesario analizar la calidad de lo disponibilizado con el fin de favorecer un adecuado estudio de resultados (información con valor agregado).

APORTES DE LA TESIS:

- Relevamiento del estado situación actual de los aspectos más relevantes en el tratamiento de datasets públicos abiertos a nivel nacional como internacional.
- Propuesta de un conjunto de métricas críticas y no críticas para analizar la calidad de datos abiertos;
- Desarrollo propio de una herramienta para validar desde varios aspectos estándares, la calidad de los datasets publicados en portales de datos abiertos estatales;
- Detección de falencias en los datasets gubernamentales disponibilizados.
- Contribuir en el análisis, verificación y comprensión del estado actual de los valores que contienen los datasets generados por las entidades gubernamentales más relevantes de la Argentina.
- Aporte sobre las mejoras en la calidad del dato y concientización de su importancia para una correcta divulgación del contenido público tanto a nivel nacional como internacional.

LÍNEAS DE I+D FUTURAS:

a) Continuar trabajando en la ampliación de la herramienta de validación HEVDA; **b)** Incorporar a la herramienta desarrollada más tipos de formatos abiertos, como ser: XML, JSON, etc. lo que llevará a establecer nuevos aspectos de control en las estructuras de los formatos que serán implementados;
c) Analizar las opciones de aplicaciones gráficas que podrían ser utilizadas con los datasets para ser embebidos en el código fuente de la herramienta HEVDA. Incorporar una herramienta de gráficos estadísticos con el análisis de las métricas críticas y no críticas; **d)** Analizar los datasets orientados a la geolocalización para el tipo de contexto: Coordenadas de longitud y latitud, formatos geoespaciales, archivos del tipo WKT (puntos de coordenadas), SHP (datos geográficos), etc.; **e)** Efectuar un estudio detallado de los datasets geoespaciales para proponer métricas de calidad de datos para este entorno de trabajo; **f)** Analizar la posibilidad de incorporar la utilización del prototipo HEVDA, en una entidad gubernamental.

CONTEXTO:

Este trabajo se enfoca en el contexto de Gobierno Abierto en aspectos del tratamiento del dato público abierto que disponibilizan las entidades estatales. Con el fin de identificar mejoras en los aspectos de calidad de los datasets abiertos, esta tesis plantea métricas críticas y no críticas para su análisis y validación de contenido, por lo que, como parte de la propuesta, se presenta un prototipo de desarrollo propio, llamado HEVDA (HErramienta de Validación de Datos Abiertos). A modo de caso de estudio, se extrae una muestra de datasets públicos estatales que son validados con HEVDA, para obtener un análisis sobre las mediciones utilizadas y realizar así, un estudio cuantitativo sobre los resultados arrojados.

Esta herramienta de validación permite detectar en forma sencilla, las falencias y errores en las fuentes de datos abiertas que podrían complicar la interoperabilidad para su utilización en diversos orígenes de bases de datos y softwares externos de otros organismos. Para evaluar la calidad de datos es necesario tener en cuenta determinadas características en el conjunto de datos analizados, por lo que se realiza un relevamiento detallado de los aspectos más notables en cuestiones de calidad de datos sobre criterios estándares de norma ISO/IEC 25012 [1], estándares universales de calidad de datos [2], dimensiones de la calidad de los datos [3], trabajos relevados y estudios realizados [4] en esta temática. En el estudio desarrollado, se puede analizar sí es factible definir métricas de calidad de datos públicos gubernamentales en un formato abierto para efectuar un análisis cuantitativo a través de una herramienta amigable y sencilla.

OBJETIVO E HIPÓTESIS:

Objetivos: El objetivo principal de la presente tesis es proponer una serie de métricas de calidad para validar los conjuntos de datos abiertos públicos gubernamentales.

Hipótesis: La hipótesis de esta tesis es la siguiente: Es posible definir métricas de calidad de datos públicos gubernamentales en un formato abierto para efectuar un análisis cuantitativo a través de una herramienta amigable y sencilla.

Reconocimiento de la tesis doctoral por parte del Ministerio de Ciencia, Tecnología e Innovación - Gobierno Argentino en el sitio web Argentina.gob.ar