Tesis Doctoral

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

Autora: Roxana Martínez

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

Doctora en Ciencias Informáticas
Facultad de Informática.
Universidad Nacional de La Plata (UNLP)
Calle 50 &, Av. 120, La Plata, Provincia de Buenos Aires, Argentina

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

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

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

RESUMEN

En estos tiempos, los datos son un recurso indispensable para cualquier actividad de gestión pública, por lo que es necesario mantenerlos actualizados, claros y completos. 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.

Palabras clave: Datos Abiertos, Datos Públicos, Gobierno Abierto, Métricas de Calidad de Datos.

1. CONTEXTO

Actualmente se utilizan las TIC (Tecnologías de la Información y la Comunicación) para la colaboración abierta en cuestiones asociadas a la resolución de diversos problemas o temas ciudadanos, que son llevados a cabo a través de sitios web gubernamentales, el enfoque principal se centra en la colaboración y participación de los ciudadanos y organismos públicos. "Esa composición la transforma en un instrumento único para impulsar la cooperación horizontal, el apoyo a la elaboración de políticas de gobierno digital centradas en el ciudadano, la formación de los funcionarios públicos, fomentar el conocimiento de aspectos claves de la construcción de una transformación digital, y el intercambio de soluciones y expertos entre los países" [5].

El Estado Nacional Argentino se encuentra promoviendo la participación ciudadana digital a través de herramientas de software en forma on-line, esta idea parte del concepto de Gobierno Abierto en la que la calidad de la información compartida por las entidades gubernamentales permite "promover y facilitar su uso siendo un instrumento que apoya el cumplimiento de los tres pilares del gobierno abierto: Transparencia, Participación y Colaboración" [6]. Si bien, el objetivo principal es la colaboración, existe otro punto a destacar que es el tratamiento de los datos colaborados, los cuales implican que, al ser compartidos, el gobierno es abierto y "permite una mayor transparencia, ofrece servicios públicos más eficientes, y alienta un mayor uso público y comercial, en la reutilización de la información del gobierno. Algunos gobiernos incluso han creado catálogos o portales (como data.gov) para facilitar al público la búsqueda y el uso de esta información" [7]. La gestión colaborativa permite la administración de los datos públicos compartidos a través de plataformas informáticas. Esta apertura de datos e información a la comunidad fortalece a la eficiencia en la gestión y optimización de las solicitudes de los ciudadanos.

El concepto de la apertura de datos públicos permite que la información de las entidades gubernamentales se presente a los ciudadanos en un formato abierto para que sea reutilizado por otros, es decir, aplicar el concepto de RISP (Reutilización de la Información del Sector Público). Si bien el Estado Nacional administra grandes cantidades de datos públicos que son propiedad de todos los ciudadanos, y puede abrir estos datos y facilitar su acceso, es importante aclarar que esto ocurrirá siempre y cuando no exponga ningún tipo de información confidencial o privada, por lo que es necesario un adecuado análisis de los conjuntos de datos que serán considerados públicos y abiertos.

Se considera que "resulta necesario aumentar la calidad de los servicios provistos por el Estado incorporando TIC (Tecnologías de la Información y de las Comunicaciones), simplificando procedimientos, propiciando reingenierías de procesos y ofreciendo al ciudadano la posibilidad de mejorar el acceso por medios electrónicos a información personalizada, coherente e integral" [8].

Según el Decreto 434/2016, la estrategia Nacional se estructura en cinco ejes: a) Plan de Tecnología y Gobierno Digital; b) Gestión Integral de los Recursos Humanos; c) Gestión por los Resultados y Compromisos Públicos; d) Gobierno Abierto e Innovación Pública; y e) Estrategia País Digital. De estos cinco ejes mencionados, existen dos que priorizan el uso de las TIC, como ser: a) Plan de Tecnología y Gobierno Digital: los cuales pretenden incorporar infraestructura tecnológica y redes que son la condición necesaria para agregar nuevos modelos de gestión basados en la tecnología de Big Data; b) Gobierno Abierto e Innovación pública: el cual no sólo se complementa con el anterior eje fundamental, sino que, además, permite la evaluación y control de las políticas públicas del Estado Nacional. Es decir que, la administración de grandes volúmenes de información es un recurso valioso y necesario. Por lo que, partiendo de esta premisa, hay una gran necesidad de gestionar esta estructura mediante la utilización e incorporación de nuevas herramientas de análisis de datos.

2. DEFINICIÓN DEL PROBLEMA

El desafío del Gobierno es brindar a los ciudadanos información de manera eficiente y transparente. En este contexto, los nuevos retos que se plantean pueden ser considerados como una oportunidad para replantear las metodologías de diseñar e implementar políticas públicas e impulsar un Estado con funcionarios colaboradores de la ciudadanía. Todo esto conlleva a la necesidad de nuevas formas de liderazgo colaborativo, y un nuevo paradigma en el tratamiento de la información abierta y pública. Este escenario, necesita promover la innovación como una "nueva actitud" y hacer frente a los nuevos desafíos de la ciudadanía digital. Es importante identificar que los pilares para la solución de esta problemática son 3 conceptos fundamentales: a) Comprender el contexto de Gobierno Abierto: apoyo del Estado Nacional y Ciudadano; b) Conjunto de datos públicos: organismos estatales deberían ofrecer una mayor cantidad de fuentes de datos de diversos temas gubernamentales con criterios preestablecidos; c) Datos abiertos: mediante la apertura de los diversos datasets en los portales de datos, utilizados como insumo fundamental de información y servicios.

Según [9] sostiene que algunos problemas que surgen en el Estado, son: a) "Procesos mal definidos"; b) "Ritmo y la magnitud del cambio estatal"; c) "Ausencia de capacidad para el aprendizaje organizacional"; d) "La resistencia del público al cambio"; e) "Ausencia de recursos": falta de apoyo financiero o de carencia de conocimientos para soporte a las iniciativas innovadoras; f) "Obstáculos técnicos": Ausencia de soluciones tecnológicas para el problema en cuestión a analizar. Es decir, es necesario implementar una serie de criterios o aspectos consensuados y sostenibles en el tiempo para mejorar la calidad de los datos públicos abiertos que se ofrecen, lo que brindará mayor poder de análisis de la información dada, como así también, nuevos escenarios sobre posibles mitigaciones de inconvenientes ciudadanos.

A modo de resumen, se puede decir que mantener la calidad y la apertura de los datos públicos, ayuda a los gobiernos y a los diferentes actores de la sociedad civil a tomar mejores decisiones, ya que tienen una visión e información de la realidad. También permite monitorear procesos, construir indicadores y hacer transparente la forma en que se gasta el dinero público [10].

Uno de los puntos a considerar en los portales de datos abiertos es que la disponibilidad de los datos no necesariamente coincide con que tengan calidad, lamentablemente, hoy sigue siendo una dificultad y es un gran desafío para las políticas públicas. El análisis de muchos de los conjuntos de datos públicos representa un problema crucial, ya que está disperso, no estandarizado y en muchos casos desactualizado. El problema en este contexto también tiene que ver con la actualización de la infraestructura técnica para los datos abiertos, ya que las prácticas de gestión de datos son débiles e inconsistentes y cambian con demasiada frecuencia. Los gobiernos deberían trabajar en la transformación técnica y organizacional de los datos abiertos, es decir, invertir en herramientas de hardware y software, como así también en estándares técnicos, capacitación, transformación organizacional y procesos de toma de decisiones para dar un soporte adecuado a la gestión de datos [11]. Como se mencionó anteriormente, aún existen algunas barreras que van desde la calidad insuficiente de los datos publicados hasta la falta de mantenimiento de los portales donde se publican [12]. Varias de estas investigaciones [13], están orientadas a la apertura de un dataset, pero no a un análisis de los problemas que puedan tener los datos abiertos que están implementados y publicados en los portales gubernamentales.

3. OBJETIVO E HIPÓTESIS

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

Objetivos específicos: Para llevar a cabo dicho objetivo se plantean los siguientes objetivos específicos: Estudiar el alcance de los conceptos implicados en este contexto de estudio; Analizar los trabajos relacionados en la temática de Gobierno Abierto relativos al tratamiento de datos abiertos y públicos; Relevar los trabajos relacionados en cuanto a los aspectos de calidad de datos abiertos y públicos; Analizar los criterios de los portales de datos abiertos para efectuar la publicación de éstos; Relevar los distintos tipos y formatos disponibles de datos abiertos que existen en los dataset actuales de los sitios más relevantes y gubernamentales de Argentina; Analizar las falencias de los datasets disponibles en los sitios de Gobierno; Establecer criterios estándar de calidad de datos; Proponer métricas críticas y no críticas de calidad de datos abiertos para validar diversos datasets; Desarrollar una herramienta de validación de datasets gubernamentales; Definir y recolectar la muestra de datasets para ser testeados por la herramienta propuesta como validadora de calidad; Definir un análisis cuantitativo en la medición de las métricas propuestas a través de la herramienta desarrollada; Analizar los resultados arrojados con la herramienta programada para las métricas de calidad propuestas; Validar las métricas sugeridas en base a criterios estándares.

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*.

4. ENFOQUE METODOLÓGICO

Para el proceso metodológico de investigación de la tesis doctoral, se define un proceso sistemático cualitativo, que implementa una forma evolutiva incremental en cada una de las etapas involucradas, siendo éstas:

1. Identificación del Problema: Definición del problema a tratar;

2. Revisión Teórica: Búsqueda, selección, recopilación, revisión y análisis del marco teórico y los trabajos relacionados pertinentes al contexto del tema de la tesis;

3. Recolección de Datos: Mediante el estudio de la revisión teórica, se definió un alcance de muestra y se extrajeron datasets de los distintos portales de datos abiertos gubernamentales en Argentina;

4. Clasificación y Análisis de Datos: Se realiza una segmentación de los datasets de los portales abiertos y se analizan los resultados en aspectos de métricas de calidad de los datos.

5. Formulación de la Solución: Se definió el alcance de la propuesta de métricas críticas y no críticas en base al estudio general realizado, identificando niveles de criticidad para éstas.

6. Aplicación de la Solución: Se desarrolló una herramienta de software para obtener de forma automática, un análisis cuantitativo para cada una de las métricas de calidad propuestas, a través del relevamiento de datasets extraídos de los portales gubernamentales de datos abiertos;

7. Validación de la Solución: Se realizó la validación de las métricas sugeridas en base a los distintos aspectos definidos que debe reunir un conjunto de métricas para la calidad de datasets abiertos. Esta definición parte del estudio realizado sobre la ISO/IEC 25012, dimensiones de calidad, el estándar universal de calidad de datos de dos capas, trabajos de estudio propio en artículos de investigación y la bibliografía relevada en esta tesis.

Cabe destacar que la estrategia / metodología formal que permite identificar la investigación es del tipo "investigación-acción", ya que este tipo de metodología se encarga del estudio de una problemática social específica, siendo en este caso el enfoque de la calidad del contenido de los datasets con información pública del estado nacional gubernamental (Información relevante a obras públicas, salud, educación, tecnologías, etc.). Este tipo de investigación propone una solución a un problema detectado, en este caso, las falencias reveladas en la calidad del contenido de los

datos públicos abiertos. En lo que respecta al planteo de una solución, esta metodología utilizada en la tesis presenta una propuesta de herramienta de desarrollo propio que permite validar los datos abiertos y, además, brindar un estudio detallado sobre lo encontrado en dicha verificación a modo de sugerencia de buenas prácticas.

5. PROPUESTA

Para analizar la calidad de datos es necesario cuantificar características en el conjunto de datos analizados. Para el estudio de la calidad de datos, esta tesis se orienta en los aspectos más relevantes en cuestiones de calidad de datos sobre criterios estándares de la 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.

La utilización de métricas de calidad favorece al encuadre de indicadores que permitan obtener un dato más limpio para facilitar el análisis final. "Los indicadores para calidad de datos son herramientas importantes que debemos tomar en cuenta en los procesos de análisis, ya que permite medir y controlar la eficiencia de nuestros procesos que derivarán en análisis y toma de decisiones dentro de una estructura organizacional" [15].

Esta propuesta de métricas de calidad se basan en el estudio de guías y buenas prácticas de publicaciones de datos abiertos gubernamentales [16], en el formato recomendado para los tipos de datos que está mayormente basado en las especificaciones de la W3C [17], en la experiencia de estudios realizados con datasets relevados de diversos sitios gubernamentales de portales de datos abiertos [4], varios criterios enfocados en la calidad del dato (Normas y estándares) y guías elaboradas por el Gobierno de la Ciudad de Buenos Aires, las cuales plantean una serie de criterios para tener en cuenta en la apertura y tratamiento del contenido de los datasets públicos. Estas [18] fueron elaboradas por el equipo de la Dirección Nacional de Datos e Información Pública junto con la Iniciativa Latinoamericana por los Datos Abiertos (ILDA), y su implementación fue posible gracias al financiamiento del fondo de transparencia del BID, que se beneficia de las generosas contribuciones de los gobiernos de Canadá, Noruega y Suecia [19]. La propuesta de métricas para el tratamiento de la calidad de los conjuntos de datos en formatos abiertos surge a raíz de un trabajo de investigación propio y publicado en la IEEE [20]. Las métricas se clasifican de la siguiente manera:

- ***Métricas Críticas:*** Contienen aquellas métricas que permiten detectar problemas de datos de una índole prioritaria para un correcto análisis de resultados con datasets, como ser: cuestiones de redundancia, contenido faltante en registros o bien datos erróneos. Es decir, es necesario tener en cuenta estos aspectos, ya que su presencia no favorece a un correcto estudio de los datos disponibilizados. Las métricas críticas propuestas: [MÉTRICA 1] Tratamiento de Números Decimales; [MÉTRICA 2] Registros Duplicados; [MÉTRICA 3] Datos Faltantes y/o Completos; [MÉTRICA 4] Caracteres inválidos.
- ***Métricas No Críticas:*** Contienen aquellas métricas que pudieran representar problemas de contenido en el dataset. Su detección está enfocada a posibles estimaciones de casos de errores y datos triviales, como así también, descubrimientos de datos redundantes combinados (entre campos y/o registros del dataset) que podrían conducir a inconvenientes en el análisis de un conjunto de datos. Las métricas no críticas propuestas: [MÉTRICA 5] Redundancia para el dominio de una columna; [MÉTRICA 6] Redundancia entre campos de una misma fila; [MÉTRICA 7] Detección de valores ID; [MÉTRICA 8] Campos Triviales.

6. VALIDACIÓN

Cada métrica propuesta, se la relacionó con un criterio de calidad que surgen del estudio de: a) Norma ISO/IEC 25012 [1]; b) Estándar Universal de Calidad de Datos [2]; c) Dimensiones de la calidad de los datos

(CDDQ) propuestas por Dan Myers [3]. El equipo de trabajo "Total Data Quality Management Program" [21], definieron un conjunto de atributos y/o dimensiones para medir y gestionar la calidad de los datos que enfocadas en la evaluación que puede ser automatizada para valorar la idoneidad y adecuación de los datos en orden a objetivos de negocio o bien necesidades. Estudios posteriores han ido modificando esta clasificación y fueron modificando diversas dimensiones. Uno de los estudios más recientes engloba algunas terminologías ya conocidas y proponen otras, como ser [3], que presenta una lista de dimensiones para la calidad de los datos, y lleva a cabo encuestas anuales con el fin de medir el uso de las dimensiones de la calidad de los datos por parte de las organizaciones.

Para este trabajo, por cada métrica propuesta, se la relacionó con una dimensión o criterio de calidad, que son el resultado de tomar en consideración, distintas fuentes:

a) Norma ISO/IEC 25012 [1], que especifica un modelo general de calidad de datos que se encuentran definidos en un formato estructurado dentro de un sistema informático. Para este se presentan los criterios del modelo de calidad de datos definido por el estándar ISO/IEC 25012, de las 15 características que lo componen, se tomaron para este trabajo, los criterios de: ***Exactitud, Completitud, Consistencia y Precisión***.

b) Estándar Universal de Calidad de Datos [2], son los criterios que debe contener un conjunto de datos para que puedan ser de calidad e interoperable y que son definidos por el estándar universal de la calidad de los datos de 2 capas. Para este aspecto se trabaja con el estándar universal en español extraído de la guía de estándares de calidad e interoperabilidad de los datos abiertos del gobierno de Colombia [22], para la validación de las métricas propuestas, se consideran los criterios de: a) Confidencialidad, siendo éstos: ***Precisión, Integridad, Consistencia y Completitud***; b) ***Presentación, siento éste el enfoque de estructura***.

c) Dimensiones de la calidad de los datos (CDDQ) propuestas por Dan Myers en DQMatters [3]. Los criterios considerados para la comparativa de validación son: ***Completo, Exactitud, Consistencia, Integridad, Precisión y Representación***.

d) Trabajos relevados y estudios de desarrollo propio para la tesis [4], [14], [20], y [23].

En base a la relación de cada criterio aplicado a las métricas propuestas, se identifica que éstas influyen en uno o más criterios de validación. el criterio. Por lo que, se observa que el criterio más representativo para validar la calidad de datos abiertos es el aspecto de Consistencia con un 92,85%, seguida por el aspecto de Integridad con un 85,71%. Luego, con un doble empate en porcentaje, el criterio de Precisión y Exactitud con el 57,14%. A éstos les sigue el criterio de Estructurales/Representación con el 50%, continuando con el 42,86% para el criterio de Relación entre valores de campos. Finalmente, concluye el criterio Redundancia con el 35,71%.

Esto conlleva a que las métricas propuestas se enfocan en un mayor porcentaje en métricas de validación orientadas a la "Consistencia", como así también a la "Integridad" de los valores que contienen los datasets, por lo que su importancia radica en la coherencia general de los datos y el valor fiable que se encuentra almacenado. Por ejemplo, un conjunto de datos que contiene información sobre las personas internadas actualmente en un hospital determinado, se la puede considerar inconsistente si el recuento de las personas es mayor que el número de camas registradas. En lo que respecta a la Integridad, reúne un conjunto de características de calidad de datos mayores, relacionando más parámetros, como ser la precisión, fiabilidad y coherencia de los datos.

7. IMPLEMENTACIÓN

Herramienta Desarrollada - HEVDA:

Como parte de la propuesta de este trabajo, se presenta una herramienta desarrollada que

permite la validación de las distintas métricas sugeridas para un conjunto de datos abiertos en formatos CSV (valores separados por comas). Si bien la herramienta HEVDA permite obtener un análisis automático, no modifica el dataset de origen, sino que brinda un estudio detallado que sirve como guía práctica orientativa para la corrección de este. Esta herramienta de desarrollo propio fue parte del trabajo de investigación elaborado y publicado en una revista internacional [14].

La funcionalidad general del software HEVDA consiste en seleccionar un archivo dataset del tipo de formato CSV y efectuar la validación de las métricas propuestas. Para la elección del tipo de formato a considerar se realizó un estudio producto del cual se publicó un artículo [4], en el que se tomó como caso de muestra el portal gubernamental Argentina Unida [24] con sus 973 datasets a julio del año 2020. Sus resultados concluyeron que el formato más utilizado es el tipo CSV con un 61,6% de uso, de allí dicha elección.

Las funcionalidades detalladas de la herramienta programada son: Identificación del cumplimiento de las métricas críticas; Detección y detalle de los casos que no cumplen con el formato válido para el tipo de dato decimal; Cálculo estimativo de los tipos de datos de los campos del dataset validado; Cálculo de la cantidad y porcentaje de los registros duplicados; Detalle de los registros duplicados; Cálculo de la cantidad y porcentaje de los registros completos; Cálculo de la cantidad de casos que poseen campos con registros Nulos (Sin Datos ni espacios en los campos); Cálculo de la cantidad de casos que poseen campos con registros Vacíos (Sin Datos y con espacios en los campos); Cálculo de la cantidad de casos que poseen campos con registros No Disponibles (Con datos que indique N/D, N/A, NULL, -, - -, --); Visualización del detalle de los casos con registros Nulos, Vacíos y No Disponibles; Cálculo de la cantidad de columnas afectadas con caracteres especiales y su correspondiente detalle; Cálculo de la cantidad de columnas afectadas con valores repetidos en un mismo campo (dominio de valores); Detalle con filtros de búsqueda para los campos del dataset y palabras de los campos detectados/as en los casos que hubo registros con valores repetidos para un mismo campo (dominio de valores); Cálculo de la cantidad y porcentaje de casos detectados con redundancia entre los valores de los campos para un mismo registro; Filtro de búsqueda para los casos detectados con datos redundantes entre los valores de los campos para un mismo registro y su correspondiente detalle; Estimación de la cantidad de ID identificados en las columnas del dataset, y la visualización de éstos; Cálculo de la cantidad de columnas afectadas con posibles campos triviales y su identificación.

Por otra parte, cabe destacar que al efectuar un análisis en un conjunto de datos se pueden encontrar una serie de aspectos que impiden el correcto análisis de validaciones, estos aspectos son de carácter bloqueante. Previo al comienzo del análisis con la herramienta HEVDA, un archivo dataset no deben cumplir con una serie de condiciones a las que se denominan “bloqueantes”, esto es, en caso de que exista algún tipo de bloqueante en la estructura del archivo, éste no podrá ser validado por el prototipo, ya que impedirá el procesamiento y análisis de las métricas propuestas del conjunto de datos. Los tipos de bloqueantes propuestos: a) El archivo posee doble caracter " (comilla); b) El archivo no cumple con la misma cantidad de columnas en cada uno de sus registros; c) El archivo no cumple con el formato CSV de separador/delimitador (coma); d) El archivo no posee una primera fila de títulos/nombres de las columnas del dataset; e) El archivo no es del tipo CSV.

8. CONTRIBUCIÓN CIENTÍFICA

Las contribuciones principales de este trabajo de tesis son las siguientes:

- 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.

9. CONCLUSIONES: REFLEXIONES FINALES

El paradigma de Gobierno Abierto, hoy por hoy desempeña un rol fundamental entre la conexión de distintos organismos gubernamentales y los ciudadanos. Gracias a este nuevo concepto es posible una mejor transparencia en base al tratamiento de rendición de cuentas por parte de entidades estatales, como así también, de una participación más comprometida con los ciudadanos de un país. Disponer de diversas guías que orienten la mejora constante de calidad de los datos abiertos es fundamental, pero, además, es vital contar con herramientas que permitan una rápida validación para tener una mejor visualización en cuanto a falencias o falta de integridad en los conjuntos de datos con el fin de aplicar las mejoras correspondientes en estos. Tener a disposición datos públicos abiertos de calidad, permitirá a los ciudadanos, una mejor confianza en las fuentes de datos y seguimiento de procesos administrativos del Estado Nacional.

Los resultados de los sondeos realizados evidencian que 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, sencilla e intuitiva para tratar aspectos de buenas prácticas.

Existen varias carencias en aspectos de calidad de los datos que se brindan en los datos abiertos, por lo que compartir buenas prácticas, guías de aprendizaje, foros o herramientas que permitan un mejor análisis y tratamiento de los datos es un beneficio para el ciudadano que desea acceder a la información pública. La propuesta de la herramienta HEVDA permite validar rápidamente el estado de "salud" de un conjunto de datos y analizar algunas de las métricas propuestas en aspectos de calidad, esto ayudará a saber si un conjunto de datos debe ser modificado o si está en condiciones de servir como base y analizar aspectos más profundos de los problemas de interoperabilidad del software para obtener valor agregado.

10. LÍNEAS FUTURAS

Algunas de las líneas futuras de investigación son las siguientes: 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.

11. PUBLICACIONES

Editorial: Editora Artemis. ebook "Ciências Socialmente Aplicáveis: Integrando Saberes e Abrindo Caminhos". Curitiva, Brasil. 2022.
Título: "Detección de errores ortográficos para la validación de la calidad en datos abiertos gubernamentales para la métrica del factor syntactic correctness".

Revista Abierta de Informática Aplicada (RAIA). 2022. Título: "Tipos de Métricas de calidad para validar datasets gubernamentales argentinos"

Journal of the School of Engineering of the Antioquia University, Colombia. 2022. Título: "Quality Study of open government data related to COVID-19 in Latin America".

Journal of Science and Research: Revista Ciencia e Investigación. 2021. Título: "Quality evaluation of government open data sets in Argentina using the HEVDA Validation Tool".

Journal: IEEE Latin America Transactions, 100(XXX). 2021. Título: "Metrics proposal to measure the quality of governmental datasets".

Congreso Nacional de Ingeniería Informática / Sistemas de Información (CoNaIISI). 2022.
Trabajo I: Título: "Análisis de la Interoperabilidad en la República Argentina. Propuesta: Prototipo para el estudio de Catálogos de Datos Abiertos".
Trabajo II: Título: "Tecnologías de la información facilitadoras para la interoperabilidad de software en gobierno abierto".

Workshop de Investigadores en Ciencias de la Computación (WICC). 2022. Título: "Propuesta de técnicas de validación para la calidad de datos abiertos e identificación de patrones para predicciones con Machine Learning".

Congreso Nacional de Ingeniería Informática / Sistemas de Información (CoNaIISI). 2021. Título: "Análisis de la Apertura de datos gubernamentales en los portales provinciales de la República Argentina: Aplicación HEVDA".

50 JAIIO. Jornadas Argentinas de Informática. 2021. Título: "Validación de Métricas propuestas de Calidad para el estudio de datos abiertos en base a criterios estándares: Aplicación HEVDA"

V Biennial Congress of IEEE Argentina Section.IEEE. UTN Argentina, Resistencia. Virtual. 2020. Título: "Analysis of datasets and catalogs in government open portals of the Argentine Republic".

8o Congreso Nacional de Ingeniería Informática – Sistemas de Información CoNaIISI. 2020. Trabajo I: Título: "Validación de la Calidad en Datos Abiertos con respecto a la detección de errores ortográficos utilizando la métrica del factor Syntactic Correctness". Trabajo II: Título: "Propuestas de Buenas Prácticas para la Implementación de Smart City en el contexto de Datos Abiertos para la Salud Pública".

14° Simposio Argentino De Informática en el Estado - SIE (49° JAIIO). 2020. Título: "Análisis del procedimiento para la solicitud de información pública y tiempos de respuesta. Caso de Estudio: Ciudad Autónoma de Buenos Aires (Argentina)".

III Congreso Internacional de Ciencias de la Computación y Sistemas de Información (CICCSI). 2019. Título: "Criterios para el análisis preliminar de datos extraídos mediante web scraping".

Congreso Argentino de Ciencias de la Computación (CACIC). 2019. Título: "Análisis de técnicas de raspado de datos en la web – aplicado al portal del Estado Nacional Argentino".

BIBLIOGRAFÍA

[1] ISO 25012 (2008), "*Ingeniería de software - Requisitos de calidad y evaluación de productos de software (SQuaRE) - Modelo de calidad de datos*", Disponible en:

https://www.iso.org/obp/ui/es/#iso:std:iso-iec:25012:ed-1:v1:en

[2] Cai, L., & Zhu, Y. (2015), "*The challenges of data quality and data quality assessment in the big data era*", Data science journal, 14.

[3] Conformated Dimensions of Data Quality (2018), "*Annual Survey about Use of Dimensions of Data Quality*", Disponible en:

http://dimensionsofdataquality.com/dims_survey

[4] Martínez, R., Rodríguez, R., & Vera, P. (2020), "*Analysis of datasets and catalogs in*

government open portals of the Argentine Republic", In 2020 IEEE Congreso Bienal de Argentina (ARGENCON) (pp. 1-8). IEEE.

[5] BID – Banco Interamericano de Desarrollo (2021), "*Gobierno Digital*", Disponible en:

https://www.iadb.org/es/dialogo-regional-de-politica/la-red-de-gobierno-electronico-de-america-latina-y-el-caribe

[6] Olaya, Y. E. L. (2018), "*Estudio sobre minería y visualización de datos abiertos del gobierno de Colombia*", Working papers, Maestría en Ingeniería de Sistemas, 2(2).

[7] W3C (2009), "*Publishing Open Government Data*", W3C Working Draft 8 September 2009, Disponible en:

https://www.w3.org/TR/gov-data/

[8] Modernización del Estado (2018), "*Serie de Investigaciones: Gobierno Abierto, Gobierno Digital y País Digital*", Instituto Nacional de la Administración Pública (INAP), Disponible en:

https://www.argentina.gob.ar/sites/default/files/inap_serie_investigaciones_empleo_publico_modernizacion-estado-2018.pdf

[9] Ramírez-Alujas, Á. V. (2010), "*Innovación en la gestión pública y open government (gobierno abierto): Una vieja nueva idea*", Innovation in Public Management and Open Government: An Old New Idea, Revista Buen Gobierno, (9)

[10] Open Contracting Partnership, Colman R. (2020), "*Women win one in four contracts in the Dominican Republic thanks to inclusive procurement reforms*", Disponible en:

https://www.open-contracting.org/2020/09/23/women-win-one-in-four-contracts-in-the-dominican-republic-thanks-to-inclusive-procurement-reforms/

[11] WebFoundation.org (2018), "*El barómetro de los datos abiertos. Edición de los líderes de la promesa al progreso*", Open Data Barometer, Disponible en:

https://webfoundation.org/docs/2018/09/WF_ODB_Report2_Spanish_Screen.pdf

[12] Cadena-Vela, S., Fuster-Guilló, A., & Mazón, J. N. (2019), "*Publicando datos abiertos considerando criterios de calidad*".

[13] Abella, A., Ortiz-de-Urbina-Criado, M., & De-Pablos-Heredero, C. (2018), "*Indicadores de calidad de datos abiertos: el caso del portal de datos abiertos de Barcelona",* El profesional de la información (EPI), 27(2), 375-382.

[14] Martínez, R., Pons, C., Rodríguez, R., & Vera, P. (2021), "*Quality evaluation of government open data sets in Argentina using the HEVDA Validation Tool*", Journal of Science and Research: Revista Ciencia e Investigación. ISSN 2528-8083, 6(2).

[15] Graph Everywhere (2021), "*Principales indicadores para Calidad de Datos*", Disponible en: https://www.grapheverywhere.com/principales-indicadores-para-calidad-de-datos/

[16] Datos.gob.ar (2021), "*Estándares según el tipo de Datos*", Disponible en:

https://datosgobar.github.io/paquete-apertura-datos/guia_abiertos/#estandares-segun-el-tipo-de-datos

[17] W3C (2015), "*Modelo para datos tabulares y metadatos en la Web*", Disponible en: https://www.w3.org/TR/tabular-data-model/

[18] Secretaría de Modernización. Presidencia de la Nación, "*Paquete de Apertura de Datos de la República Argentina*", Disponible en: https://datosgobar.github.io/paquete-apertura-datos/guia-subnacionales/#1-que-son-los-datos-abiertos

[19] Datos.gob.ar (2021), "*¿Por qué es importante estandarizarlos*", Disponible en: https://datosgobar.github.io/paquete-apertura-datos/guia-interoperables/#por-que-es-importante-estandarizarlos

[20] Martínez, R., Rodríguez, R. A., & Vera, P. M. (2021), "*Metrics proposal to measure the quality of governmental datasets*", IEEE Latin America Transactions, 100(XXX), Disponible en: https://latamt.ieeer9.org/index.php/transactions/article/view/5642

[21] The MIT Total Data Quality Management Program (2002), "*MIT TDQM Program Highlight*", Disponible en:

http://web.mit.edu/tdqm/www/index.shtml

[22] Gobierno de Colombia (2020), "*Guía de datos, LEILA – Librería de calidad de datos*", Unidad de Científicos de Datos Dirección de Desarrollo Digital, Disponible en:

https://colaboracion.dnp.gov.co/CDT/Desarrollo%20Digital/Big%20Data/2020/00_LEILA/LEILA_Presentacion.pdf

[23] Martínez, R., et al. (2020), "*Validación de la calidad en Datos Abiertos con respecto a la detección de errores ortográficos utilizando la métrica del factor Syntactic Correctness*", Congreso Nacional de Ingeniería Informática y Sistemas de la Información, Universidad Nacional de Tecnología (UTN), Facultad Regional San Francisco.

[24] Argentina unida, "*Datos Argentina*", Disponible en: https://datos.gob.ar/