Metrics Proposal to Measure the Quality of Governmental Datasets

María Roxana Martínez, Rocío Andrea Rodríguez and Pablo Martín Vera

Abstract— **The government challenge is to provide citizens with information efficiently and transparently. In this context, the new challenges that arise can be considered as an opportunity to rethink the methodologies of designing and implementing public policies and promoting a State with collaborating public officials. All this leads to a new paradigm in the treatment of open and public information. Currently, data is an indispensable resource for any public management activity, so it is necessary to keep it updated and complete. Although it can be determined that more and more governments are embarking on this new concept of open data, there is still a long way to go. Several related works on this subject are increasingly focused on the quality of open data from the portals of government entities, but it is still necessary to reinforce and understand the importance of the data, that is, if a data has quality, it can be better used, manipulated and distributed to citizens for various purposes. This work focuses on the proposal of quality metrics to analyze the contents of published datasets and thus obtain an analysis to improve their dissemination and interoperability between databases and software.**

Index Terms— **Best practices, Government, Open Access.**

I. Introducción

La gestión de los servicios otorgados por el Estado Nacional, debe darse a conocer mediante la participación activa de los ciudadanos en la evaluación y el control de los diversos programas del Estado y de las instituciones públicas, con el fin de mantener la confianza y un vínculo activo entre los intereses de los pobladores. El gobierno abierto se enfoca en la "construcción de estados transparentes, participativos, que rindan cuentas de manera adecuada e innovadores, poniendo al ciudadano en el centro de la toma de las decisiones públicas, como una forma de fortalecer el Estado democrático. El gobierno abierto se basa en cuatro pilares fundamentales: transparencia, rendición de cuentas, participación ciudadana y colaboración e innovación pública y ciudadana" [1]. Con la ayuda de los datos abiertos públicos que pueden ser consultados por cualquier persona u organismo de forma libre [2], se logra el acceso que es la precondición al ejercicio de otros derechos y a la fuente de ciudadanía política [3].

M. R. Martínez, profesora e investigadora, Universidad Abierta Interamericana, Buenos Aires, Argentina (e-mail: roxana.martinez@uai.edu.ar).

R. A. Rodríguez, profesora e investigadora, Universidad Abierta Interamericana, Buenos Aires, Argentina (e-mail: rocioandrea.rodriguez@uai.edu.ar).

P. M. Vera, profesor e investigador, Universidad Abierta Interamericana, Buenos Aires, Argentina (e-mail: pablomartin.vera@uai.edu.ar).

Los datos que se pueden reutilizar y redistribuir sin ninguna restricción se denominan datos abiertos [4]. Cada vez son más los portales estatales que proporcionan datos públicos ya sea para usuarios finales ó para alimentar otros sistemas. Por lo que, para cada conjunto de datos públicos es necesario determinar su valor, su calidad y quién es el responsable de su mantenimiento/actualización. Es por esto que, los avances tecnológicos para el análisis de grandes bases de datos, datos estadísticos y datos abiertos pasan a tener un rol fundamental, creando posibilidades sin precedentes para informar a la sociedad [5].

La transparencia puede ser de dos tipos: pasiva o activa. La transparencia activa consiste en que los organismos públicos brinden un libre acceso a la información a través de su portal web, en un formato abierto. La transparencia pasiva se asocia con el derecho del acceso a la información que produce el Estado. Este artículo tiene un enfoque en la transparencia activa, ya que se orienta en la calidad de los datos abiertos que son brindados a través de los portales abiertos de distintos organismos estatales, en las siguientes secciones se analizan las características de éstos y posteriormente, diversos aspectos a tener en cuenta en dicho contexto.

II. Planteamiento del Problema

El marco teórico-conceptual del que surgen los argumentos por los cuales las administraciones públicas se están reordenando y rediseñando para asumir ciertos rasgos de la empresa privada moderna, consisten en que se pueden formar equipos de trabajo del conocimiento competitivos que fomenten las TIC (tecnologías de la información y comunicación) como vínculos entre la sociedad y sus administraciones públicas, por lo que, las metodologías para evaluar la relación y el impacto de las TIC en las administraciones públicas actuales aún son incipientes y continúan desarrollándose [6]. La gestión pública actual, requiere de un cambio interno en las pautas culturales, organizacionales, normativas y en su relación con el ciudadano. La agilidad, la innovación y la capacidad de adaptación a cambios constantes en el entorno se convirtieron en nuevos valores públicos. La innovación podría mejorar las actividades del gobierno con las necesidades de los ciudadanos actuales, y ayudar a resolver algunos inconvenientes que se presentan en el diseño e implementación de las políticas públicas de hoy. Uno de los enfoques que proponen los gobiernos, es disponibilizar conjuntos de datos públicos en portales de sitios web gubernamentales en formato abierto, con el fin de difundir la transparencia ciudadana y reforzar el vínculo entre Estado y ciudadanos. Si bien este paradigma se encuentra establecido en

varios países del mundo, aún es necesaria la incorporación de datasets en varios organismos que aún no han incursionado en dicho tema. Uno de los grandes problemas encontrados en estos portales de datos abiertos, es la inconsistencia de algunos de los datasets, esto se debe a que no todos los conjuntos son disponibilizados bajo un determinado estándar, sino que los organismos publican en base a guías de sugerencias y/o un criterio predefinido por ellos mismos (en donde diferentes organismos podrían tener distintos criterios). No existe un estándar internacional definido obligatorio en el que puedan respaldarse y trabajar en común.

Según estudios [7], [8], dentro del contexto tecnológico existe una gran preocupación por las problemáticas relacionadas con los datos, básicamente aspectos como la interpretación e intercambio, la disgregación de bases de datos, la falta de interoperabilidad, como así también, la ausencia de modelos estándares de datos son graves inconvenientes a los que se enfrentan muchos países en sus portales gubernamentales. Es necesario implementar una serie de criterios o aspectos consensuados entre los organismos gubernamentales y que éstos sean 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 a inconvenientes ciudadanos. Algunos aspectos [9] a tener en cuenta en dicho tratamiento es el análisis e implicancia de nuevas ideas, lograr responder a una necesidad específica para los ciudadanos, como así también contemplar un cambio en los procesos que se vienen llevando a cabo para el desempeño institucional. Éstos generan resultados observables que luego contribuyen al fortalecimiento público, lo que fomentaría la participación de las personas y/o entidades estatales de todo el mundo en el concepto de transparencia.

III. Trabajos Relacionados

Las implementaciones de los datos abiertos en la política pública, pueden ser analizados desde el punto de vista de una mejor usabilidad en los portales de datos abiertos, es decir, sitios abiertos más amigables para los ciudadanos. Para ello, algunas de las propuestas [10] [11], sugieren incorporar un análisis de información detallado, que puedan ayudar en el proceso de inculcar la cultura de gobierno abierto a las personas. En [12] se propone un prototipo de software para la evaluación de principios de datos abiertos que permitan, además, validar el cumplimiento de leyes de los datos abiertos, como estudios recientes que se realizaron en los portales abiertos de Colombia, a partir de los principios establecidos por el Open Goverment Data [13]. Por lo que, analizar los principios de los datos abiertos y las leyes que involucran políticas sobre el acceso a la información, es importante, para la implementación de una arquitectura que pueda evaluar los datos recolectados de plataformas de datos abiertos para establecer la madurez de este tipo de información. Otra propuesta [14] se enfoca en la evaluación de portales de datos abiertos en base a criterios definidos por los autores, como ser el planteamiento de metodologías e indicadores que miden la calidad de éstos para los sitios web de datos abiertos en los países de España, Brasil, Costa Rica, Taiwán y la Unión Europea. De este estudio se presentan dos enfoques bien marcados: (a) Datos publicados: abarcando la calidad (identifican factores de disponibilidad, actualización, accesibilidad, visualización, formatos de publicación y completitud), uso (identifican los factores de demanda definida, número de visualizaciones, descarga, consumo de API y productos resultantes) y metadatos (identifican el factor de uso, completitud y Recuperabilidad); (b) Portales de datos abiertos: resaltando los aspectos de su estructura (identifican el factor de categorización), usabilidad (identifican el factor de búsqueda, Navegabilidad y uso/consumo/descarga de datos) y mecanismos de comunicación (identifican los factores de comentarios y discusión, fuente-usuarios y solicitud).

Cada metodología aporta un enfoque distinto en la medida en que están planteados sus criterios de evaluación, lo cual puede llevar a que el elemento estudiado (portal o datos) tenga niveles de calidad distintos, según la metodología que se haya empleado. Por otro lado, otro trabajo [15] realiza foco en nuevos modelos de medición del nivel de uso de datos abiertos del gobierno abierto, con técnicas de mediciones de aceptación, uso y grado de confianza de los usuarios. Varias investigaciones, utilizan los informes publicados por el sitio oficial del barómetro de datos abiertos reconocido internacionalmente [16], con el fin de medir el éxito y el impacto de las iniciativas de datos abiertos de los gobiernos en todo el mundo. Al medir la intención conductual de usar tecnologías, los investigadores adaptan teorías y modelos relacionados con ese objetivo. En cuanto a la evaluación de la usabilidad [17], se realizaron propuestas de buenas prácticas, en base a la investigación de la utilización de datos para los países de Australia, Canadá, India, Estados Unidos y Reino Unido, para mejorar la capacidad de las partes interesadas para descubrir, acceder y reutilizar estas fuentes de información en línea. Existen trabajos [18], [19], [20], [21], [22], [23] que se enfocan en establecer criterios y clasificaciones de niveles de calidad de los datos abiertos. Algunos escenarios de análisis se orientan a partir de la reutilización de datos abiertos y públicos.

IV. Solución Propuesta

Actualmente existen diversos conjuntos de datos (datasets) que son brindados por organismos gubernamentales, tanto a nivel nacional como internacional en sus portales de datos abiertos. Muchas veces, estos datos no poseen calidad, es decir, no están en un formato abierto, existen caracteres inválidos, o bien se encuentran campos incompletos entre otros problemas. La propuesta de este trabajo es generar métricas que permitan analizar su calidad. Considerando que los datos son un recurso indispensable para cualquier actividad de gestión pública, por lo que es necesario mantenerlos actualizados y completos. Si bien se puede determinar que cada vez son más los gobiernos que se embarcan en este nuevo paradigma de concepto de datos abiertos, todavía falta un largo camino por recorrer necesitando una base de métricas unificadas que permitan evaluarlos.

A. Calidad de Datasets Gubernamentales

La calidad de los datos favorece a una mejor utilización,

manipulación y distribución a los ciudadanos para varios fines, permitiendo a su vez, una mayor participación ciudadana y transparencia en organismos estatales. "La reutilización de datos abiertos permite el desarrollo de nuevos productos y servicios digitales, creando oportunidades de desarrollo social y económico. Sin embargo, la reutilización de estos datos se enfrenta con diversas barreras en su expansión, debido a diversos problemas relacionados con la calidad de datos que van desde la incompletitud de los datos hasta la de actualización de los mismos. En concreto, un criterio relevante de calidad de datos abiertos es la comprensibilidad, ya que un proceso de interpretación errónea provocará ambigüedades o malentendidos que desmotivará su reutilización" [24]. Para medir la calidad de datos es necesario cuantificar determinadas características en el conjunto de datos analizados. Para el estudio de la calidad de datos, este trabajo se orienta en algunos aspectos de la ISO/IEC 25012 [25]. 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 nuestros procesos de análisis ya que nos permite medir y controlar la eficiencia de nuestros procesos que derivarán en análisis y toma de decisiones dentro de una estructura organizacional" [26].

Esta propuesta de métricas se basa sobre las buenas prácticas de publicación de los datos abiertos [27], y en el formato recomendado para los distintos tipos de datos que está mayormente basado en las especificaciones de la W3C [28], y, sobre todo, en la experiencia de datasets relevados de diversos sitios gubernamentales de portales de datos abiertos [29]. En las siguientes secciones se presentan las métricas propuestas para el tratamiento de la calidad de los conjuntos de datos en formatos abiertos. Las métricas se clasifican en métricas críticas y no críticas. Previamente, se realiza una breve explicación de los conceptos de dimensiones utilizadas.

B. Dimensiones de Calidad Utilizadas

La dimensión de calidad de datos, unicidad [33], indica el nivel de duplicación de los datos. Esto ocurre cuando un objeto del mundo real se encuentra representado más de una vez en los datos, esto es, varias tuplas (registros) representan exactamente el mismo objeto. Por lo que la duplicación es cuando la misma entidad aparece repetida de manera exacta. Básicamente, la unicidad [34], mide el grado en que un dato está libre de redundancias en amplitud, profundidad y alcance. En amplitud las propiedades y clases representadas, En alcance, una base de conocimiento en donde múltiples ejemplares representan el mismo objeto; En profundidad donde múltiples valores de una propiedad son únicos. Es decir que, "las medidas de unicidad permiten entender que los valores distintos de un elemento de datos aparecerán una sola vez reflejados en el conjunto de datos" [26]. Por otra parte, para la dimensión de calidad de datos, completitud [33], indica si el sistema de información contiene todos los datos de interés, y si los mismos cuentan con el alcance y profundidad que sea requerido. Además, indica que existen dos factores de la completitud: cobertura y densidad: a) La cobertura se refiere a la porción de datos de la realidad que se encuentran contenidos en el sistema de información; b) La densidad se refiere a la cantidad de información contenida y faltante acerca de las entidades del sistema de información.

La dimensión exactitud es el grado en el que los datos representan correctamente el verdadero valor del atributo deseado de un concepto o evento en un contexto de uso específico. Tiene dos principales aspectos de los cuales se considerará la "Exactitud Sintáctica", es decir, la cercanía de los valores de los datos a un conjunto de valores definidos en un dominio considerado sintácticamente correcto [25]. Esto puede darse por un carácter inválido/especial.

C. Métricas Críticas

Son 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 propuestas en esta categoría son: (1) Números Decimales: Esta métrica analiza los datos del tipo decimal, verificando si cumplen con las siguientes recomendaciones: Valores numéricos con 2 dígitos decimales; Formato de separación de decimal con "." (punto); y Valores numéricos sin separador de mil. (2) Registros Duplicados: Esta métrica propuesta permite detectar y mostrar los registros que se encuentran duplicados en el conjunto de datos analizado. Uno de los puntos fundamentales a tener en cuenta en las fuentes de datos abiertos gubernamentales, es la correcta utilización de estos, es por ello que detectar los casos de duplicación de registros favorece en una mejor manipulación de los mismos. Esto a su vez, permite definir estructuras de datos y simplicidad en el tratamiento de los distintos procesos que se podrían utilizar con ellos, por ejemplo, procesos de Extracción, Transformación y Carga, ETL (Extract, Transform and Load) que brindan a las organizaciones/empresas/organismos una gestión de datos desde múltiples fuentes. Algunos de los proyectos en gestión pública [31] [32], procesan datos abiertos para detectar variables que predigan incumplimientos, o bien se desarrollaron modelos estructurales, proyectos de simulaciones, o bien se utilizan los datos como parte de un pronóstico para relevar inspecciones de manera más eficaz. Como menciona el autor Beltrán [30] en relación a los datos sucios como un gran inconveniente, indica que "suelen tener problemas que afectan significativamente a la precisión del modelo predictivo y pueden llevar a conclusiones erróneas, por ejemplo, pueden no contener suficiente información, pueden no estar elegidos al azar o, simplemente, pueden ser erróneos". Como indica que autor Alba Cuellar [35], la correcta "detección de registros duplicados en archivos electrónicos que contienen información acerca de objetos del mundo real […], es un proceso de importancia fundamental si se desea generar, analizar y diseminar información estadística de buena calidad". Uno de los enfoques que garantiza la calidad en los datos, se relaciona con el aspecto de dimensiones, para esta métrica se propone la dimensión de Unicidad, esta se encuentra en el factor de calidad: No-duplicación en el cual se estudia el grado de duplicación de la fuente de datos analizada. Dentro de esta, se trabaja en varios aspectos, por ejemplo, para esta métrica puntual, en la redundancia de los registros de un conjunto de datos. (3) Datos Faltantes y/o Completos: Esta métrica

propuesta se divide en el análisis de dos partes. Como primera parte, permite detectar la cantidad de registros en los que todos sus campos están completos, y el porcentaje que representa sobre el total de los registros del conjunto de datos. Como segunda parte, se detectan datos faltantes analizándolos desde distintas perspectivas, como ser: que el dato no esté ya que se presenta en forma vaciá, o que se completa con algún texto particular que indica que el dato no se encuentra, por ejemplo: espacios, guiones, N/D, etc. La importancia de tener todos los valores cargados en sus correspondientes campos, permite un correcto análisis de manera completa sobre los datos, como así también lograr tener información para una adecuada interpretación, y así efectuar un estudio en detalle de lo analizado. Básicamente, la falta de valores en los campos del dataset, proporciona una delgada línea a la confusión y/o error de interpretación de casos, ya que muchos de estos datos abiertos, son utilizados en tablas dinámicas, algoritmos estadísticos, historias de datos abiertos [36], visualizaciones de gráficos o bien desarrollos de software. Al igual que la métrica de detección de casos de los Registros Duplicados de la sección anterior, para el análisis de esta métrica, se toman en cuenta medidas cuantitativas de calidad de datos. Para este enfoque que garantiza la calidad en los datos, se orienta el aspecto de dimensiones al concepto de: Completitud. "El nivel de completitud de datos refleja el grado en el que todos los atributos de un dato están presentes, lo que permite tener una visión clara sobre la integridad de los elementos a estudiar" [26]. En un modelo relacional, la completitud (en particular, la densidad) se caracteriza principalmente por los valores nulos, cuyo significado a pesar de ser variado, es importante conocer. Uno de los temas importantes a tener en cuenta con la dimensión completitud, se debe a que uno de los mayores problemas, se da por valores nulos o vacíos. Esto conlleva a pensar que un valor puede reportarse como nulo, porque tal vez se omitió dicha información por parte de la fuente de datos, o bien por un error de almacenamiento en la base de datos y/o sistema que lo generó o tal vez no se conoce. En base a esto, es que se considera necesario, identificar estos casos para un análisis más detallado de éstos. Desde la guía de sugerencias de buenas prácticas para la publicación de datos en formatos abiertos, se sugiere el tratamiento e identificación de los valores nulos, desconocidos o en blanco en un conjunto de datos, y que los elementos o celdas en blanco se interpretarán siempre como "valor ausente" [37]. Por este motivo, este trabajo propone 3 grupos de casos para darle tratamiento a este tipo de aspectos: Nulos, Vacíos y No Disponibles (explicados anteriormente). Otra de las recomendaciones que indica el sitio de datos abiertos de Argentina [39], es "no dejar celdas vacías en filas bajo la presunción de que valores en blanco posteriores a un valor positivo contienen implícitamente a ese mismo valor en una suerte de agrupamiento conceptual. Este error es muy común en la construcción de planillas de cálculo y suele generar problemas graves cuando cambia el orden original de las filas. Además, impide el uso de tablas dinámicas y otras formas de analizar los datos" [38], es decir, se sugiere indicar un valor redundante en caso de que corresponda relación lógica, antes que dejar un campo vacío. Esto es con el fin de facilitar el análisis de datos en un agrupamiento lógico conceptual. (4) Caracteres Inválidos: Esta métrica propuesta permite identificar los caracteres especiales del conjunto de datos analizado. Se puede incluir el carácter afectado, y el número de registro del dataset, como así también, el nombre de la columna/campo en el que aparece. Es importante localizar este tipo de caracteres con el fin de no alterar la identificación y análisis de los valores contenidos en los conjuntos de datos. El gran problema que conlleva que los datos se presenten con caracteres inválidos, supondrá una pérdida de información y, por consiguiente, una pérdida de la objetividad de lo que se está analizando como resultado. Al igual que la métrica de detección de datos faltantes y/o completos de la sección anterior, para el análisis de esta métrica, se toman en cuenta medidas cuantitativas de calidad de datos. Para este aspecto, la dimensión de calidad de datos utilizada es Exactitud.

D. 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 en esta categoría son: (1) Redundancia para el dominio de una columna: Esta métrica propuesta permite calcular la cantidad de valores repetidos en una misma columna/campo, es decir, trabaja con el dominio de datos (conjunto de valores posibles). Para analizar este aspecto, se muestra la cantidad de columnas afectadas que poseen valores iguales entre si sobre la cantidad total de registros del conjunto de datos. Dentro de los estándares de calidad, recomendado por el sitio gubernamental de datos abiertos de la República Argentina [39], se sugiere que las entidades que aparezcan entre los datos de un campo textual deben tener una descripción única. Por ello, la importancia de detectar casos de valores igual, con el fin de saber si están bien agregados o deben ser modificados para que cumplan con una misma descripción. Por lo que se sugiere que toda mención que se realice a una entidad dada debe hacerse usando exactamente la misma cadena de caracteres cada vez [27]: a) Las descripciones de entidades deberían elegirse siempre de forma tal que cumplan con el estándar específico que las describe, en caso de que este exista; b) Cuando este estándar no existe y hay dudas respecto del criterio a adoptar para elegir la descripción única de una entidad, debe privilegiarse siempre aquella que sea lo más explícita, descriptiva y declarativa posible. Un ejemplo: para el caso de cuatro valores de texto que refieren a la misma entidad, siendo: Ciudad Autónoma de Buenos Aires; CABA; Capital Federal; Ciudad de Buenos Aires, lo que se podrían deducir como una descripción estándar recomendada por la guía de buenas prácticas [41], es "Ciudad Autónoma de Buenos Aires". Este presenta un agrupamiento lógico definido, que permitirá un mejor análisis a futuro sobre la información establecida. Es decir, siempre que sea posible, es importante incorporar un estándar para identificar un marco común en el tratamiento de un aspecto dado. (2) Redundancia entre campos de una misma fila: Esta métrica propuesta permite identificar la cantidad de casos con redundancia entre los valores de los campos para un mismo registro. Un principio importante para la calidad de datos es no repetir los mismos valores en más de una columna para la misma fila del dataset. Esto es, debe estar bien definida y descripta cada columna que se brinda en el conjunto de datos,

ya que representa un valor específico en el análisis lógico y representativo de los datos en formato abierto. Uno de los principales problemas a nivel bases de datos, es la detección y tratamiento de datos duplicados, es decir, encontrar varios registros exactamente iguales en una tabla, ya que esto trae aparejados problemas de diseño e inconsistencias. Para mejorar la calidad de los datos, es necesario eliminar la información redundante o repetitiva. La duplicación de los datos puede conllevar a equivocaciones o bien errores lógicos en el análisis final que pueden ser consecuencia de no contar con un enfoque integrado en la lógica del dataset. Como buena práctica, el sitio gubernamental de datos abiertos de la República Argentina [39], sugiere como una de las pautas a tener en cuenta en la construcción de los datasets es que "no se deben repetir nombres entre los campos" [46]. Esta métrica se relaciona con la dimensión de Unicidad. Para esta dimensión de calidad, también se utiliza el factor de calidad: No-duplicación en el cual se analiza la cantidad de casos con redundancia entre los valores de los campos para un mismo registro. (3) Detección de valores ID: Esta métrica permite realizar una estimación sobre la cantidad de campos "ID" identificados en las columnas del dataset. Para ello, se analizan los nombres que contengan, tanto en mayúsculas como en minúsculas, los siguientes elementos: 'id'; 'id_'; '_id'. En los campos que contienen en sus descripciones "ID", son utilizados con valores numéricos y representan un código o valor único de enteros que no es nulo, y, además, se implementan para identificar de forma unívoca a cada una de las filas del conjunto de datos. Resulta necesario cuantificar la cantidad de campos para este tipo, ya que, si bien son identificadores, a los ciudadanos y/u organismos que acceden a este conjunto de datos, podrían no interpretar el significado del código numérico que se muestra, esto es, en muchos casos los datasets forman parte de un análisis de diversos estudios estadísticos, que no logran detectar el significado y/o utilidad de la nomenclatura tradicional de códigos ID. Esto se resolvería, si en el sitio oficial del cual fue extraído el dataset, se acompaña el diccionario de datos correspondiente, con el fin de comprender los campos del conjunto de datos, y, sobre todo, el significado de estos campos ID. En algunos datasets de Argentina, el ID es el identificador único del catálogo dentro de la Red de Nodos de Datos Abiertos de la Administración Pública Nacional. Este identificador es otorgado por la Dirección Nacional de Datos e Información Pública cuando un nuevo nodo pide ser incorporado a la red para su federación en el nodo concentrador de datos abiertos de la APN (Administración Pública Nacional) [45], como buena práctica se sugiere que sea una o más palabras en minúsculas, separadas con guiones medios, sin usar caracteres especiales. Cabe aclarar que no todos los conjuntos de datos tienen campos ID, esto es de forma opcional. Si bien los datasets abiertos, deben contener datos bien definidos, organizados y justificados, como buena práctica, el sitio gubernamental de datos abiertos de la República Argentina [39], sugiere la utilización de un campo identificador en el dataset, debido a que "suele ser útil para la identificación unívoca de variables en algunos sistemas o aplicaciones, pero no en la generalidad de los casos" [40]. Cabe aclarar que no todos los conjuntos de datos tienen campos ID, esto es de forma opcional. Desde el portal de datos abiertos de la República Argentina, se presenta una guía orientativa, que busca ayudar a los organismos a instrumentar la Política de Datos Abiertos impulsada desde el Gobierno de la Nación Argentina, a través del Decreto N° 117/2016 del 12 de enero de 2016. "Esta es una guía de buenas prácticas para el uso de entidades interoperables. Se trata de datos básicos y fundamentales cuyo uso se repite frecuentemente entre datasets de temáticas y fuentes distintas. Para hacer estas recomendaciones, nos basamos en estándares usados a nivel nacional e internacional y en la experiencia de trabajo del equipo de la Dirección Nacional de Datos e Información Pública del Secretaría de Gobierno de Modernización de la Jefatura de Gabinete de Ministros de la Nación. Esta es una guía colaborativa y en progreso. Valoramos, y alentamos, a organizaciones y ciudadanos a plantear ideas, sugerencias, y comentarios que nos ayuden a crear un mejor documento" [41]. Es importante destacar que los datasets incluyen campos de ID que son los que permiten que el conjunto de datos sea interoperable entre varios datasets. "Las entidades interoperables son las que permiten que los datasets hablen entre sí, pero esto no puede suceder cuando dos datasets nombran de forma distinta a una misma entidad interoperable (como cuando se usan distintos sistemas de id o se nombra una misma entidad con/sin mayúsculas, usando artículos y preposiciones (o no usándolos), usando abreviaturas, siglas, tildes, forma corta o completa de un nombre, etc. Para que los datasets puedan ser interoperables, deben identificarse todas las entidades interoperables presentes en un dataset y asegurarse de que los datos sobre ellas siguen el mismo estándar" [42]. (4) Campos Triviales: Esta métrica propuesta permite identificar si existen campos con valores iguales para una misma columna. Esta métrica se relaciona con la dimensión de Unicidad, utiliza el factor de calidad: duplicación en el cual se analiza el grado de repetición de valores de un mismo campo. Un registro duplicado ocurre cuando el mismo dato se ha introducido más de una vez, por lo que es importante, detectar si existen campos/columnas que poseen en todos sus valores el mismo dato. El descubrimiento de estos casos permitirá conocer si hay campos que pueden ser omitidos en el dataset, ya que éstos podrían indicarse como dato en el nombre del conjunto de datos. Por ejemplo: si en todos los registros se detectara un campo país = Argentina, entonces el dataset debería contener en su nombre Argentina, siendo: Dataset llamado "Casos registrados de Covid-19", podría llamarse "Casos registrados de Covid-19 en Argentina".

La solución propuesta podría ser implementada en los portales de datos abiertos gubernamentales, en secciones que están orientadas a las guías de buenas prácticas para la apertura de los datos abiertos públicos [2], esto permitiría que los organismos que necesiten disponibilizar datasets, tengan la posibilidad de acceder a una herramienta de validación de la calidad de los datos, antes de ser publicados, como así también, dar acceso a los ciudadadanos que deseen utilizar la aplicación para analizar datasets ya divulgados y conocer su estado.

V. Validación y Resultados

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 [25], que especifica un modelo general de calidad de datos que se encuentran definidos en un formato estructurado

dentro de un sistema informático. (b) Estándar Universal de Calidad de Datos [43], 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. (c) Dimensiones de la calidad de los datos (CDDQ) propuestas por Dan Myers en DQMatters [44]. (d) Trabajos relevados. (e) Estudios realizados en datasets abiertos de portales gubernamentales de la República Argentina.

En la Tabla 1 con el resumen del análisis, identificando que todas las métricas propuestas influyen en uno o más criterios escogidos (esto se indica con una X). Se puede observar que el criterio más representativo es el aspecto de Consistencia (Cr.5) con un 87,50% (es decir, de las 8 métricas, se cumple en 7), que es el grado en el que los datos están libres de contradicción y son coherentes con otros datos en un contexto de uso específico [25]. Los siguientes aspectos son: Integridad (Cr.4) y Precisión (Cr.1) con un 75% (es decir, de las 8 métricas, se cumplen en 6), siendo que la integridad enfoca la calidad estructural de los datasets, y se relaciona con la validez, duplicación y coherencia de éstos [45], y, por otro lado, la precisión que abarca el detalle de la medición que se utiliza para especificar un determinado dominio para un campo, identificando así, el grado en el que los datos tienen valores que son exactos o proporcionan discernimiento [25]. Como criterios menos relavantes, se observan los Estructurales (Cr.7) y Completitud (Cr.3) con el 12,50% (es decir, de las 8 métricas, sólo se cumple 1) en los que se mide el grado en el que todos los datos se encuentran completos [25] e intervienen aspectos de diseño estructural del dataset. Esto muestra como se cobrieron todos los criterios por medio de las métricas diseñadas.

TABLA 1
CRITERIOS DE VALIDACIÓN

Nro. Métrica	Cr. 1	Cr. 2	Cr. 3	Cr. 4	Cr. 5	Cr. 6	Cr. 7	Cr. 8
1	X				X		X	
2		X		X	X	X		X
3			X					
4	X	X		X	X			
5	X	X		X	X	X		X
6	X	X		X	X	X		X
7	X			X	X			
8	X	X		X	X	X		X

En las columnas se muestran los distintos criterios de validación de las métricas propuestas. Se indica con una X si influye la métrica en el criterio. Los criterios utilizados son: Cr.1=Precisión; Cr.2=Exactitud; Cr.3=Completitud; Cr.4= Integridad; Cr.5= Consistencia; Cr.6= Relación entre valores; Cr.7=Estructurales/Representación; Cr.8=Redundancia.

VI. CONCLUSION

Según se muestra en este estudio, el papel que desempeñan las TICs es fundamental para la gestión de los datos abiertos públicos, por lo que es necesario verificar y realizar un seguimiento constante de todos los conjuntos de datos que son compartidos a través de los portales de sitios web gubernamentales, ya que son las fuentes de distribución para generar nuevo conocimiento como valor agregado a la comunidad. La mayoría de estos datos suelen disponibilizarse en formatos estandarizados como, por ejemplo: CSV, XLS, XLSX, XML, entre otros. La posibilidad de optar por varios formatos, facilitará que un tercero pueda elegir el más conveniente para visualizarlos o bien utilizarlos como entrada de otra herramienta o sistema. Disponer de diversas guías como las métricas propuestas en este trabajo, facilita el enfoque en la mejora de la calidad de los datos abiertos, ya que es fundamental para establecer una correcta interoperabilidad entre las tecnologías que los utilizan. Esta propuesta permite sugerir un punto de vista para tener una rápida validación de los temas centrales en este contexto y así, lograr una visualización panorámica de un dataset en cuanto a falencias o falta de integridad en los conjuntos de datos con el fin de aplicar las correcciones correspondientes en estos. Esta investigación, propuso métricas orientadas a la calidad para facilitar dichas mediciones, estableciendo criterios críticos y no críticos en el análisis, lo que permitir clasificar un relevamiento de los aspectos que son vitales a considerar en una primera instancia, y luego observar los criterios que podrían llegar a conducir a contingencias o problemas en el análisis de resultados con datasets. Aplicando las métricas explicadas, se obtiene un rápido estudio sobre el estado situación de un dataset, es decir, en caso de que se identifiquen errores de la índole propuesta, puede ayudar a los organismos en el proceso de interoperabilidad de los datos consumidos en los posibles sistemas informáticos que los utilicen. Tener a disposición datos públicos abiertos de calidad permitirá a los ciudadanos, una mejor confianza en la información brindada, como así también un seguimiento de procesos administrativos del Estado Nacional.

REFERENCIAS

[1] Arroyo Chacón, J. (2017), “La Innovación Abierta Como Pilar Del Gobierno Abierto”, *Open Innovation as a Pillar of Open Government, Revista Enfoques*, 15(27), pp. 13-41.

[2] 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, consultado abril 2020.

[3] Manfredi-Sánchez, J. L. (2017), “Horizontes de la información pública. El *profesional de la información* (EPI)”, 26(3), pp. 353-360.

[4] E. Oviedo, JN Mazón y JJ Zubcoff (2013), "Hacia un modelo de calidad de datos para portales de datos abiertos", *XXXIX Latin American Computing Conference* (CLEI), Naiguata, 2013, pp. 1-8.

[5] Montero, Gregorio (2017), “Del gobierno abierto al Estado abierto: la mirada del Centro Latinoamericano de Administración para el Desarrollo. Desde el gobierno abierto al Estado abierto en América Latina y el Caribe”, Santiago: CEPAL. LC/PUB. 2017/9-P. pp. 53-81, 2017.

[6] Ávila Barrios, D. (2014), “El uso de las TICs en el entorno de la nueva gestión pública mexicana”, Andamios, 11(24), pp.263-288.

[7] Jiménez, C. E., Criado, J. I., & Gascó, M. (2011). Technological e-government interoperability. an analysis of iberoamerican countries. *IEEE Latin America Transactions*, 9(7), pp.1112-1117.

[8] Penteado, B., Carlos, M. J., & Isotani, S. (2021). Process model with quality control for the production of high quality linked open government data. *IEEE Latin America Transactions*, 19(3), pp. 421-429.

[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] Rodríguez, J. A. M. (2019), “Valoración de factores de uso de los datos abiertos de gobierno”, Instituto de Ciencias de Gobierno y Desarrollo

Estratégico (Doctoral Dissertation, Benemérita Universidad Autónoma De Puebla).

[11] Paños V., A., & Jordán-Alfonso, A. (2017), "Acceso A La Información Pública Y Su Reutilización En Las Comunidades Autónomas: Evaluación De La Reutilización De Datos Abiertos", El profesional de la información, 26(3).

[12] Ariza A. D. F., & Rojas Clavijo, J. A. (2019), "Prototipo de Software para la evaluación de principios de datos abiertos". Universidad Católica De Colombia, Facultad De Ingeniería, Programa De Ingeniería De Sistema, Trabajo De Investigación Tecnológica, Bogotá D.C., Colombia.

[13] OECD Better policities for better lives, "Open Government Data".

[14] Melo, C. A. H., & Sanabria, J. S. G. (2020), "Proposal for the Evaluation of Open Data Portals", Facultad de Ingeniería, 29(54), pp. 1-20.

[15] Zainal, N. Z., Hussin, H., & Nazri, M. N. M. (2019), "Acceptance, Quality and Trust Factors–Conceptual Model for Open Government Data Potential Use", International Journal on Perceptive and Cognitive Computing.

[16] Open Data Barometer – World Wide Web Foundation, "The Open Data Barometer", Disponible es: https://opendatabarometer.org/?_year=2017&indicator=ODB, consultado en marzo 2020.

[17] Máchová, R., Hub, M., & Lnenicka, M. (2018), "Usability evaluation of open data portals", Aslib Journal of Information Management.

[18] Oviedo Blanco, E. (2016), "Modelo de madurez para portales de datos abiertos e incorporación a la norma técnica nacional de Costa Rica", Repositorio Institucional de la Universidad de Alicante, Tesis Doctoral.

[19] Beltrán, L., Estefan, N., & Mahecha Moyano, J. F. (2017), "Prototipo de software para la evaluación de la calidad de datos abiertos", Tesis de grado, Repositorio Institucional de la Universidad Católica de Colombia.

[20] Ibanez Gonzalez, L., Millard, I., Glaser, H., & Simperl, E. (2019), "An assessment of adoption and quality of linked data in European open government data".

[21] Rodríguez Rojas, L. A. (2017), "Metamodelo para integración de datos abiertos aplicado a inteligencia de negocios", Tesis de Doctoral, Repositorio Institucional de la Universidad de Oviedo.

[22] Arizo, I. (2016), "Métricas basadas en datos", Tesis de Maestría en Gestión de la Información, Universitat Politècnica de València.

[23] 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).

[24] Cadena-Vela, S. (2019), "Marco de referencia para la publicación de datos abiertos comprensibles basado en estándares de calidad", Tesis doctoral en Ciencias Informáticas, Universidad de Alicante.

[25] ISO 25012 (2008), "Ingeniería de software - Requisitos de calidad y evaluación de productos de software (SQuaRE) - Modelo de calidad de datos".

[26] Graph Everywhere (2021), "Principales indicadores para Calidad de Datos".

[27] 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, consultado en enero 2021.

[28] W3C (2015), "Modelo para datos tabulares y metadatos en la Web", Disponible en: https://www.w3.org/TR/tabular-data-model/, consultado en enero 2021.

[29] Roxana Martínez, Rocío Rodríguez, Pablo Vera (2020), "Análisis de datasets y catálogos en los portales abiertos gubernamentales de la República Argentina", *IEEE ARGENCON 2020. V Biennial Congress of IEEE Argentina Sectio*n, In virtual mode, December 1 to December 4, 2020.

[30] Beltrán Martínez, B. (2014), "Minería de datos", Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación.

[31] Gobernarte Ideas innovadoras para mejores gobiernos (2017), "Análisis predictivo: Impulsar mejoras mediante el uso de datos"

[32] Gómez, C. E. J., & Roma, J. C. (2018), "Análisis predictivo de datos abiertos sobre el uso turístico del servicio de alquiler compartido de bicicletas de Nueva York", Universidad Oberta de Catalunya, Master Universitario en Ciencia de Datos.

[33] Marotta, A., Vallespir, D., & Valverde, C. (2012), "Análisis de la calidad de datos en experimentos en ingeniería de software", In XVIII Congreso Argentino de Ciencias de la Computación.

[34] Méndez Matamoros, J. H. (2017), "Mejoramiento de calidad en conjuntos de datos abiertos basado en la aplicación de métricas de consistencia lógica", Tesis de Maestría en Ciencias de la Información y las Comunicaciones.

[35] Alba Cuellar, D. (2011), "Detección de registros duplicados entre dos archivos digitales".

[36] Buenos Aires DATA (2021), "Historias con Datos", Ciudad de Buenos Aires, Disponible en: https://data.buenosaires.gob.ar/historias-con-datos, consultado en marzo 2021.

[37] Datos.gob.ar (2021), "Valores nulos, desconocidos o en blanco en campos numéricos", Disponible en: https://datosgobar.github.io/paquete-apertura-datos/guia_abiertos/#valores-nulos-desconocidos-o-en-blanco-en-campos-numericos, consultado en enero 2021.

[38] Datos.gob.ar (2021), "Celdas vacías en filas para agrupar conceptos", Disponible en: https://datosgobar.github.io/paquete-apertura-datos/guia_abiertos/#celdas-vacias-en-filas-para-agrupar-conceptos, consultado en enero 2021.

[39] Argentina unida, "Datos Argentina", Disponible en: https://datos.gob.ar/, consultado marzo 2020.

[40] Perfil de Aplicación Nacional de Metadatos para Datos Abiertos, Secretaría de Modernización, Presidencia de la Nación, "Perfil de Aplicación Nacional de Metadatos para Datos Abiertos".

[41] Datos.gob.ar (2021), "Guía para la identificación y uso de entidades interoperables", Paquete de Apertura de Datos de la República Argentina

[42] Datos.gob.ar (2021), "¿Porqué es importante estandarizarlos", Disponible en: https://datosgobar.github.io/paquete-apertura-datos/guia-interoperables/#por-que-es-importante-estandarizarlos, consultado en enero 202.

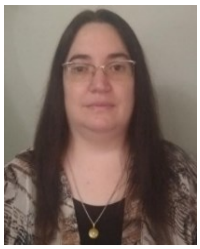
[43] Cai, L., & Zhu, Y. (2015), "The challenges of data quality and data quality assessment in the big data era", Data science journal, 14.

[44] Conformated Dimensions of Data Quality (2018), "Annual Survey about Use of Dimensions of Data Quality".

[45] Datos Argentina – Paquete-apertura-datos (2017), "Guía para el uso y la publicación de metadatos", Disponible en: https://paquete-apertura-datos.readthedocs.io/es/0.2.3/guia_metadatos.html, consultado en marzo 2021.

[46] Datos.gob.ar (2021), "CSV", Guía para la publicación de datos en formatos abiertos, Disponible en: https://datosgobar.github.io/paquete-apertura-datos/guia_abiertos/#csv, consultado en enero 2021.

María Roxana Martínez, Argentina, Ingeniera en Sistemas Informáticos (UAI-Universidad Abierta Interamericana). Doctorando en Ciencias Informáticas en la Universidad Nacional de La Plata (UNLP). Magíster en Tecnología Informática (UAI). Docente de posgrado en UAI. Docente de grado en UAI, UdeMM (Universidad de la Marina Mercante) y UNQ (Universidad Nacional de Quilmes). Contenidista en la Universidad Siglo 21. Investigadora en UAI. Es autora de artículos en congresos y revistas. Ha participado como tutora y jurado de tesis grado y posgrado, revisora en congresos nacionales. En el ámbito laboral, posee más de 18 años de experiencia en la rama de IT, actualmente se desempeña como Líder de Procesos IT en la UIF (Unidad de Información Financiera) de Argentina.



Rocío Andrea Rodríguez, Argentina, Ingeniera en Informática (UNLaM - Universidad Nacional de La Matanza). Doctora en Ciencias Informáticas (UNLP - Universidad Nacional de La Plata). Docente de grado en UNLaM, UTN (Universidad Tecnológica Nacional) y UAI (Universidad Abierta Interamericana). Docente de posgrado en UAI. Desde el 2005 realiza investigación académica, actualmente es directora de proyectos en UAI). Tiene categoría 2 en el Programa de Incentivos al Docente investigador. Dirige tesis de grado, maestría y doctorado. Ha sido jurado en tribunales de tesis, revisora de artículos científicos en congresos y revistas. Autora de diversas publicaciones en congresos, revistas y libros.

Pablo Martín Vera, Argentino, Ingeniero en Informática recibido en la Universidad Nacional de La Matanza (UNLaM). Obtuvo su título de Doctor en Ciencias Informáticas en la Universidad Nacional de La Plata (UNLP). Actualmente es docente de grado y postgrado en UNLaM, Universidad Tecnológica Nacional (UTN) y en la Universidad Abierta Interamericana (UAI). Adicionalmente a la docencia, se desarrolla como director de proectos de investigación en UAI. Cuenta con mas de 100 publicaciones académicas. Es revisor de trabajps científicos en diferentes congresos y revistas.