HACIA LA EVALUACION DE LA CALIDAD DE DATOS ABIERTOS

Ana Funes[(1)], Aristides Dasso[(1)], María Alejandra Barrera[(2)]

[(1)]Departamento de Informática, Facultad de Ciencias Físico-Matemáticas y Naturales, Universidad Nacional de San Luis
Ejército de los Andes 950 - 5700 San Luis, Argentina
afunes@unsl.edu.ar, arisdas@unsl.edu.ar

[(2)]Departamento de Informática, Facultad de Tecnología y Ciencias Aplicadas, Universidad Nacional de Catamarca,
Maximio Victoria 55, Catamarca, Argentina
mbarrera@tecno.unca.edu.ar

RESUMEN

Según la Open Knowledge Foundation[1], los datos abiertos son datos que están disponibles de forma pública a cualquier persona que desee utilizar, reutilizar y redistribuir libremente, sujetos únicamente, como máximo, al requisito de atribuir y compartir por igual.

Estos datos suelen estar disponibles en formatos digitales y se publican bajo diversos tipos de licencias. Considerando los beneficios derivados de su adopción, múltiples iniciativas han sido creadas con el objetivo de promover su uso, tales como la Alianza para el Gobierno Abierto (Open Government Partnership), el Portal de Datos Abiertos de la Unión Europea y el Portal de Datos Abiertos de los Estados Unidos, entre otras. En Argentina, la iniciativa Datos Argentina[2] del gobierno argentino tiene como objetivo mejorar la transparencia y la participación ciudadana mediante la publicación de datos abiertos.

Frente a este escenario, consideramos importante contar con modelos y herramientas automatizadas que permitan llevar a cabo evaluaciones de la calidad de repositorios de datos abiertos, y que sirvan de guía a los organismos gubernamentales al evaluar la calidad de los datos que publican, contribuyendo de esta forma en la transparencia y la rendición de cuentas, así como en la resolución de problemas sociales y económicos importantes.

***Palabras clave*:** datos abiertos, calidad, métodos de evaluación multicriterio, métricas.

CONTEXTO

El presente trabajo se encuentra enmarcado en una investigación conjunta entre investigadores del Proyecto de Ciencia y Técnica PROICO 03-2020 ″Ingeniería de Software: Estrategias de Desarrollo, Mantenimiento y Migración de Sistemas en la Nube″, de la Universidad Nacional de San Luis. (Director: Daniel Riesco. Acreditado con evaluación externa. Financiamiento: Universidad Nacional de San Luis) e investigadores del Proyecto: ″ Ingeniería de Software en la era de las Industrias 4.0.″, que se desarrolla en el departamento de Informática de la Facultad de Tecnología y Ciencias Aplicadas de la Universidad Nacional de Catamarca, también acreditado con evaluación externa.

En este contexto, se ha venido trabajando desde hace tiempo en el ámbito del SEG (Software Engineering Group) de la Universidad Nacional de San Luis, sobre la construcción de modelos de evaluación de sistemas complejos, donde se han obtenido resultados que han sido plasmados en diversas publicaciones (ver por ejemplo [5], [6], [7], [9], [10], [11], [12]).

[1] https://opendatahandbook.org/guide/en/what-is-open-data/
[2] https://www.datos.gob.ar/

1. INTRODUCCIÓN

Según la Open Knowledge Foundation[3] el conocimiento es abierto si cualquiera es libre de acceder a él, usarlo, modificarlo y compartirlo, sujeto, como máximo, a medidas que preserven la procedencia y la apertura. Es un concepto general que se puede aplicar a diversos elementos o partes del conocimiento transferido, como por ejemplo al código abierto, a publicaciones, textos y, en particular, a los datos.

Los datos abiertos, entendidos como una de las formas de conocimiento abierto, son datos que cualquier persona puede utilizar, reutilizar y redistribuir libremente, sujetos únicamente, como máximo, al requisito de atribuir y compartir por igual [13].

Uno de los usos más inmediatos está asociado al gobierno electrónico ya que para los organismos gubernamentales es particularmente importante, no solo por la cantidad y la centralidad de los datos que recopila, sino también por el tipo de información que brinda la mayoría de esos datos: datos públicos por ley, los cuales podrían estar disponibles públicamente para su uso.

Esto contribuye a mejorar la transparencia de este tipo de organismos ya que permite a los ciudadanos monitorearlos y evaluarlos a través de la información que se hace pública.

En algunos casos, los datos abiertos también pueden involucrar a los ciudadanos en el proceso de toma de decisiones, lo que puede aumentar la participación y la colaboración entre las partes interesadas, así como mejorar la calidad de las decisiones tomadas.

En este sentido, no solo los organismos gubernamentales pueden beneficiarse de la adopción de datos abiertos. Otros organismos tanto públicos como privados también pueden verse favorecidos con estos y otros aspectos tales como la aplicación en la investigación y análisis de datos, o en la innovación y desarrollo económico.

[3] https://okfn.org/

Considerando los beneficios antes mencionados, creemos necesario disponer de métricas que permitan evaluar aspectos de calidad de los datos abiertos ya este conocimiento no solo va a contribuir, desde el punto de vista de los proveedores de datos, en mejorar la calidad de los datos publicados, ya que ayudará a identificar errores y deficiencias, sino también, desde el punto de vista del consumidor de los datos, para ganar en confianza sobre los mismos.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Considerando los beneficios que surgen del uso de datos abiertos y de la posibilidad de evaluar su calidad, nos planteamos como objetivo principal de esta línea de trabajo la creación de un modelo de atributos de calidad de datos abiertos así como la definición, sobre la base de los atributos identificados en dicho modelo, de métricas que sirvan para obtener un indicador de la calidad de los mismos.

Para lograr dicho objetivo, los ejes principales de la presente línea de investigación involucran las siguientes actividades:

Análisis bibliográfico de propuestas alternativas para la identificación y conceptualización de los atributos de calidad de datos abiertos, analizando y definiendo las propiedades específicas que deben ser tenidas en cuenta en dicho contexto.

Modelización, estableciendo métodos de cálculo o procedimientos en forma de métricas, para obtener un valor numérico por cada atributo a ser considerado.

Aplicación en la evaluación de la calidad de datos abiertos, para obtener el valor de un indicador global del Sistema/Producto evaluado.

3. RESULTADOS OBTENIDOS/ ESPERADOS

La primera de todas las actividades llevadas a cabo se basó en el análisis de

publicaciones relacionadas a la evaluación de datos abiertos, donde hemos buscado identificar los diversos atributos de calidad que han sido considerados.

Así, por ejemplo, el trabajo de A. Vetro et al. [14] presenta un marco para la evaluación de datos abiertos, el cual viene aplicado a repositorios de datos publicados por el gobierno italiano. En este marco se identifican atributos o características de calidad tales como la fiabilidad, precisión, actualidad y accesibilidad, entre un total de siete, para los cuales se definen siete métricas simples de conteo o porcentajes. No se describe ninguna métrica que integre los resultados en un indicador global.

En otro trabajo, Tim Berners-Lee publica, en 2006, un esquema de implementación para datos abiertos, basado en cinco requisitos incrementales en exigencia [4]. Según él, el conjunto de datos abierto debe cumplir con los siguientes requisitos: (1). Disponibilidad en la web, cualquier formato proporcionado con licencia abierta; (2) Disponibilidad como datos estructurados legibles por máquina (por ejemplo, Excel en lugar de escaneo de imagen); (3). Formato no propietario disponible (p. ej., CSV en lugar de Excel); (4) Hacer uso de estándares abiertos del W3C (RDF y SPARQL) y URI para identificar cosas; (5) Vinculación de los datos a los datos de otros proveedores para proporcionar contexto. Esta propuesta, considera solo un aspecto de la calidad de los datos relativa al formato o codificación de los mismos, dejando de lado aspectos de calidad tales como la integridad, la consistencia y la puntualidad.

Por otro lado, el trabajo de Arbello et al. [1] se centra en la definición de una métrica, que han dado en llamar MELODA, para la evaluación de la reutilización de datos abiertos. La misma se basa en 4 dimensiones: estándares técnicos, acceso, legal y modelo, las cuales hacen referencia al formato de los datos; al acceso o mecanismo por el cual se hace posible la descarga o conexión con la información; a las restricciones legales de acceso y a la descripción de la estructura de los datos, respectivamente.

Los mismos autores en [2] redimensionan su modelo, proponiendo una nueva métrica, MELODA5, la cual se basa en ocho dimensiones, que son una evolución de MELODA 4.13 y donde el nombre de la dimensión "modelo" es renombrada a "estandarización" y donde se agregan dos nuevas dimensiones, una para medir la diseminación de las actividades y otra para analizar la reputación de la fuente de datos.

Por otro lado, la Open Data Barometer (ODB) en la 4ta edición de su Informe Global[4], establece que para que los datos abiertos sean de valor y usables, deben, además, ser comprensibles, precisos y de alta calidad. Asimismo, establece que los proveedores de datos, deberían asegurar que cuentan con mecanismos que permitan a los usuarios proveer retroalimentación, haciendo posible revisiones continuas de sus datos con el objetivo de mejora.

Por otro lado, en el portal de la W3C[5], se brinda una serie de recomendaciones a la hora de publicar datos gubernamentales abiertos a partir de las cuales se pueden identificar atributos que hacen a la calidad de tales repositorios. También, la Sunshine Fundation[6] identifica, en su portal, una lista de diez principios que brindan una perspectiva para evaluar hasta qué punto los datos gubernamentales son abiertos y accesibles para el público.

En base a este análisis es que, en una primera etapa, hemos podido concluir que la calidad de los datos abiertos no se restringe a evaluar sólo la calidad técnica de los datos o solo aquellos aspectos propios de datos abiertos sino que, si bien consideramos importante identificar claramente aquellas dimensiones que definen a los datos como abiertos, tales como la licencia, las formas de

[4] ttps://opendatabarometer.org/doc/4thEdition/ODB-4thEdition-GlobalReport-ES.pdf
[5] https://www.w3.org/TR/gov-data/
[6] https://sunlightfoundation.com/policy/documents/ten-open-data-principles/

acceso, la legibilidad y el formato, entre otras, ya que al asegurar que se cumplen estos requisitos, se garantiza que los datos sean accesibles y utilizables por cualquier persona o entidad interesada en ellos, consideramos también importante, tal como establece la iniciativa de datos abiertos del Gobierno de España[7], no dejar de lado la calidad técnica de los mismos, que aunque no esté directamente contemplada en los principios de los datos abiertos, deberían también tenerse en cuenta a la hora de producir cualquier tipo de datos como, por ejemplo, algunos atributos de calidad definidos por la ISO/IEC 25012, que podrían adaptarse fácilmente:

Exactitud: se encuentran dentro del rango de valores válidos definidos para el dominio de aplicación.
Completitud: se aportan los valores correspondientes a todos los atributos disponibles.
Credibilidad: tanto para los datos en sí como para la fuente de información.
Actualidad: proporcionados en el momento preciso para mantener su valor.
Accesibilidad: facilidad de acceso en su contexto.
Conformidad: con respecto a los estándares y normativas vigentes.
Confidencialidad: respetando la privacidad y seguridad de los datos.
Eficiencia: para que puedan ser procesados con unos recursos razonables.
Precisión: respecto al contexto al que pertenecen.
Trazabilidad: respecto a la fuente u origen de los datos.
Comprensibilidad: con una codificación adecuado para su posterior interpretación.

Muchos de estos atributos de calidad de los datos han sido identificados también en el proyecto de datos abiertos de la Comisión Europea [8].

Finalmente, en base al análisis realizado, se pudo observar también que muchas de las propuestas analizadas consideran atributos de calidad comunes aunque bajo diferentes nombres.

En base a estas conclusiones, se planea (a) desarrollar un modelo jerárquico de atributos de calidad de datos abiertos, que unifique e incluya las dimensiones, subdimensiones y atributos necesarios para definir una métrica de calidad que considere estos aspectos antes analizados. (b) A partir del modelo jerárquico de calidad construido, definir un conjunto de métricas que permitan medir los atributos hojas de dicha jerarquía. (c) Definir una función multicriterio que agregue las características antes identificadas y que permita la evaluación tanto integral como de subcaracterísticas del modelo jerárquico. Este modelo cuantitativo de evaluación, permitirá obtener los valores de indicadores parciales de las características a diferentes grados de abstracción del modelo jerárquico de calidad, así como un valor de un indicador global del repositorio de datos evaluado. (d) Aplicar el modelo y las métricas asociadas para evaluar la calidad de los datos a diversos portales de datos abiertos.

4. FORMACIÓN DE RECURSOS HUMANOS

La línea de investigación sobre construcción de modelos de evaluación de sistemas complejos ha producido dos tesis de posgrado en la Maestría en Ingeniería de Software de la Universidad Nacional de San Luis. Una de ellas sobre evaluación de atributos de calidad de sitios de gobierno electrónico [1][6] y la otra sobre el atributo de calidad Accesibilidad en aplicaciones web [9][10]. Asimismo, uno de los autores de este trabajo ha realizado también su tesis de Maestría en Ingeniería de Software de la Universidad Nacional de San Luis en temas relacionados a la evaluación de la reusabilidad de datos espaciales abiertos [3].

La propuesta aquí expuesta, también, tiene como objetivo ser motivo de desarrollo de tesis en el ámbito del Departamento de Informática de la Universidad Nacional de San Luis, como así también en ámbitos del Departamento de Informática de la Universidad Nacional de Catamarca.

[7] https://datos.gob.es/es/noticia/datos-abiertos-y-de-calidad

5. BIBLIOGRAFIA

[1] Alberto Abella García, Marta Ortiz de Urbina Criado, Carmen de Pablos Heredero. El profesional de la información, ISSN 1699-2407, Vol. 23, Nº 6, 2014 , págs. 582-588.

[2] Alberto Abella; Marta Ortiz de Urbina Criado; Carmen De Pablos Heredero. El profesional de la información, ISSN 1699-2407, Vol. 28, Nº 6, 2019.

[3] María Alejandra Barrera. Una estrategia de evaluación de la reusabilidad de los conjuntos de datos de los portales de Infraestructuras de Datos Espaciales Tesis de Maestría en Ingeniería de Software. Universidad Nacional de San Luis, Argentina. 2022.

[4] Berners-Lee, T. (2006) Linked data-design issues. Tech. rep., W3C, http://www.w3.org/DesignIssues/LinkedData.html.

[5] M. Castro. "Análisis de las propiedades y atributos propios de sitios de gobierno electrónico", Tesis de Maestría en Ingeniería de Software, Departamento de Informática, Universidad Nacional de San Luis, 2010.

[6] M. Castro, A. Dasso, A. Funes. "Modelo de Evaluación para Sitios de Gobierno Electrónico", Simposio de Informática en el Estado (SIE) 2009 – 38 JAIIO, Mar del Plata, Argentina, August 26-28, 2009. pp. 200-214.

[7] A. Dasso y A. Funes, "Threat and Risk Assessment Using Continuous Logic", Encyclopedia of Organizational Knowledge, Administration, and Technologies, 1st. edition. IGI Global. 2020.

[8] Makx Dekkers, Nikolaos Loutas, Michiel De Keyzer and Stijn Goedertier. Open Data & Metadata Quality. European Commission. Open Data Support. 2014. https://joinup.ec.europa.eu/sites/default/files/document/2015-05/d2.1.2_training_module_2.2_open_data_quality_v1.00_en.pdf.

[9] C. Gallardo, A. Funes, H. Ahumada. "Soporte para la Medición y Evaluación de la Accesibilidad al Contenido en Aplicaciones Web", Anales de ASSE 2019 (JAIIO 2019), Salta, Argentina. pp. 56-70.

[10] C. Gallardo, A. Funes. "Un Modelo para la Evaluación de la Calidad de la Accesibilidad al Contenido Web", CONAIISI 2015, Bs. As., Argentina.

[11] E. Miranda et al. "NESSy: A new evaluator for software development tool". 2nd Symposium on Languages, Applications and Technologies. 2013.

[12] M. Peralta, C. Salgado. "Un a Herramienta para la Evaluación de Sistemas", Tesis de Licenciatura en Ciencias de la Computación, Universidad Nacional de San Luis. 2004.

[13] The Open Knowledge Foundation. The Open Data Handbook. https://opendatahandbook.org/guide/en/

[14] Antonio Vetrò, Lorenzo Canova, Marco Torchiano, Camilo Orozco Minotas, Raimondo Iemma, Federico Morando, Open data quality measurement framework: Definition and application to Open Government Data, Government Information Quarterly, Volume 33, Issue 2, 2016, Pags 325-337, ISSN 0740-624X, https://doi.org/10.1016/j.giq.2016.02.001.