Hacia la optimización del uso de datos abiertos en el ámbito público

Carrizo Claudio[*], Pablo Vaca[*], Salgado Carlos[+], Sánchez Alberto[+], Peralta Mario[+]
[*]Facultad Regional San Francisco – Facultad Regional Córdoba
Universidad Tecnológica Nacional
Av. de la Universidad 501 - San Francisco - Córdoba - Tel. 03564-421147
{cjcarrizo77, vacapablo72}@gmail.com
[+] Departamento de Informática Facultad de Ciencias Físico-Matemáticas y Naturales Universidad Nacional de San Luis
Ejército de los Andes 950 – C.P. 5700 – San Luis – Argentina
e-mail: {csalgado, alfanego, mperalta}@unsl.edu.ar

RESUMEN

En la actualidad, existe una fuerte iniciativa por parte de los gobiernos en poner los datos públicos a disposición de los ciudadanos, con el fin de que los mismos puedan aportar un valor a la información, en pos de mejorar la calidad de vida de los habitantes, impulsando lo que se denomina "Ciudades Inteligentes".

Pero desde la experiencia de los ciudadanos que consumen estos datos**,** existen dificultades al realizar procesos ETL (Extraction, Transformation and Load, en inglés) sobre todo, en lo que respecta a normalización, formatos de presentación e interpretación de los datos.

El objetivo de esta investigación consiste en proponer un modelo y algunos formatos de datos que sean apropiados al dominio del problema. El modelo de datos debe permitir caracterizar y normalizar los datos, mientras que los formatos recomendables junto con la exposición de una metadata asociada, deben permitir optimizar la presentación e interpretación de los datos abiertos u OD (Open Data, en inglés).

Este trabajo está enfocado en los portales web de datos abiertos del ámbito público, en pos de que estos puedan brindar datos que sean óptimos para ser utilizados en iniciativas innovadoras, que puedan contribuir a mejorar la calidad de vida de los ciudadanos.

Palabras clave:

Datos Abiertos, Ciudades Inteligentes, Modelo de datos, Representación de datos, Metadata.

CONTEXTO

La presente línea de investigación se encuentra inmersa dentro del Proyecto de Investigación y Desarrollo Inter-facultad "Estado del Arte en Ciencia de Datos y Big Data", el cual se encuentra homologado y financiado por la Secretaría de Ciencia, Tecnología y Posgrado de la Universidad Tecnológica Nacional, bajo el código IFN4567 y según la Disposición SCTyP Nº 468/16. El periodo de ejecución de dicho proyecto es desde el 1 Enero de 2017 hasta el 31 de Diciembre de 2018, y el mismo está incluido en el Programa I&D + i de Tecnología de las Organizaciones de la Universidad Tecnológica Nacional.

1. INTRODUCCIÓN

Con el advenimiento de las nuevas tecnologías de la información y comunicaciones (TICs) [1] [2], existe una tendencia a nivel mundial acerca de la apertura de datos desde distintas instituciones, en especial las que pertenecen al estado nacional en todos sus

niveles. En los últimos años, los gobiernos a nivel de provincias/estados y países del mundo, han avanzado en la publicación de OD [3] [4] [7], no solo como un medio para generar transparencia, sino también para alentar su uso en iniciativas innovadoras que busquen mejorar la calidad de vida de los habitantes en las ciudades, para de esta poder contribuir en la búsqueda de lo que se denomina "Ciudades Inteligentes" [5] [6]. Una ciudad es considerada inteligente cuando aplica las TICs con el fin de dotar de una infraestructura que mejore la calidad de vida de sus habitantes y sea sostenible en el tiempo.

Según un trabajo publicado en la Academia de Estudios Económicos de Bucarest [7], OD es el concepto o la idea de aquellos que piensan que los datos deben poder ser accedidos por cualquier usuario y republicados todas las veces que se quiera, sin restricción de ningún tipo. Según Maximiliano Bron en el libro "Open Data, Miradas y perspectivas de los datos abiertos" [3], hablar de OD es mucho más que una creencia o un concepto, es una práctica o una filosofía que establece que los datos deben estar disponibles sin restricciones de acceso a ellos y que es algo similar a lo que ocurre con el software libre o de código abierto.

Las iniciativas OD son movilizadas por varios motivos, según un trabajo presentado por Calvin Chan en la Conferencia Internacional sobre Ciencias de Sistemas en el 2013 [4], hay dos causales principales para abrir la información, uno de ellos es la democracia y la libertad de la información, buscando ser más abiertos y transparentes, y el otro es económico, buscando que iniciativas privadas agreguen valor a la información.

Uno de los puntos que nos motivaron para trabajar en esta temática, fue el análisis de la forma con que los distintos actores de la sociedad se apropian de los conocimientos. Esto se sustenta con la experiencia que tienen los distintos integrantes del grupo y la información recabada con los distintos actores de la empresa del medio con los que se interactúa por razones laborales, capacitaciones y estudio de posgrado. Otro de los aspectos que motivan el desarrollo de esta línea de investigación, está relacionado con las dificultades constantes que encuentran los alumnos al hacer uso de OD en el desarrollo de trabajos prácticos en la Cátedra de Big Data [8]: Arquitecturas y Estrategias de Análisis de Datos Masivos de la carrera de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Regional Córdoba. El desarrollo de estos trabajos consiste en buscar fuentes de datos abiertas en portales web en el ámbito público, en donde a través de procesos de ETL [9], se puedan obtener OD que luego serán utilizados para generar información de valor, a través de la aplicación de técnicas de Big Data o Minería de datos, en pos de lograr impacto positivo en las sociedades, mejorando de esta manera la calidad de vida de los ciudadanos. Del análisis de las dificultades encontradas por los alumnos en el uso de OD, se evidenciaron problemas de normalización, formatos de presentación e interpretación del significado de los datos abiertos.

Una posible solución al problema detectado anteriormente consiste en:

- Utilizar modelos que permitan caracterizar y normalizar OD.
- Identificar formatos de presentación según el origen y contexto de los datos
- Exponer una metadata asociada que permita interpretar el significado de los OD.

El objetivo de este trabajo consiste en realizar una discusión acerca del uso de modelos, formatos recomendables de presentación y exponer una metadata asociada para datos abiertos en el ámbito público.

Para lograr esto será necesario:

- Identificar y describir modelos que permitan caracterizar y normalizar los datos abiertos
- Analizar y evaluar distintos portales web que brinden datos abiertos públicos para identificar formatos de presentación recomendables y exposición de metadata.

Con este trabajo se propone realizar un aporte hacia los portales web de datos abiertos en el ámbito público, en pos de que los ciudadanos puedan hacer un mejor uso de los OD y de esta manera, puedan aportar valor a la información a través de sus iniciativas, contribuyendo a mejorar la calidad de vida de los habitantes en las ciudades.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Los principales ejes de trabajo de esta línea de investigación se detallan a continuación:

- Identificar y describir modelos que permitan caracterizar y normalizar datos
- Analizar y evaluar portales de datos abiertos del ámbito público a nivel nacional e internacional
- Identificar formatos de presentación de datos según el origen y contexto de los datos
- Identificar formatos de exposición de metadata que permitan interpretar el significado de los OD
- Realizar una discusión acerca del uso de modelos y el formato de presentación de OD.

3. RESULTADOS OBTENIDOS/ESPERADOS

Hasta el momento, se ha logrado identificar y describir los modelos más utilizados, entre ellos, Modelo Relacional [10], Clave-Valor, Documental [11] [12]), los cuales permitirán caracterizar y normalizar datos en pos de evitar inconsistencias y redundancias e incrementar la velocidad de respuesta sobre las consultas u operaciones que se realizan sobre ellos.

Para lograr una mejor interpretación del significado de los datos, será necesario considerar el uso de una metadata [13]. Se puede tomar como referencia la plataforma Twitter [14] [15], dicha red social permite, mediante la tecnología REST (Representational State Transfer, en inglés), utilizar la información almacenada mediante el formato JSON [16], el cual no solo contiene los valores de los campos, sino también el nombre de los mismos. Otra alternativa presentada en el trabajo "Open Metadata Formats: Efficient XML-Based Communication for Heterogeneous Distributed Systems" [17], consiste en utilizar documentos XML como metadata, de tal manera que describan los datos expuestos, tal como el nombre, tipo de dato, extensión, significado o dominio. De la revisión del trabajo de Huamin Wang y Zhiwei Ye [18], se puede observar la utilidad de contar con una metadata. En dicho trabajo, se plantea un servicio de ETL basado en metadata, la misma es utilizada para especificar las estructuras de datos y además las fuentes de datos, reglas de procesamiento o relaciones entre ellos. De esta idea, surge la importancia de que al publicar datos abiertos, estos cuenten con una metadata asociada, lo cual facilita la utilización de los mismos y asegura su correcto entendimiento.

Lo anterior se ve reflejado en el estudio y análisis de distintos portales a nivel nacional e internacional de datos en ámbitos públicos. Comparando, por ejemplo, el portal de Aragón Open Data [19] y el de CABA [20], se observa que en ambos se utilizan los formatos estándares y los exponen con una metadata mínima. De la comparación entre ambos, se puede decir que el primero contiene repositorios más formateados, ya que todos los conjuntos de datos están en los formatos XML y JSON. En el segundo, no todos los conjuntos de datos son presentados con los mismos formatos, lo que agrega un nivel de dificultad extra al momento de utilizarlos.

Del estudio realizado sobre distintos portales se ha podido determinar que:

- En ambientes con información estructurada, se encontró cierta dificultad al momento de entender el significado de la información de los campos. Este tipo de formato debe ser mejorado agregando información sobre los datos, por ejemplo a través de la metadata de los mismos, utilizando por ejemplo un formato XML o similar que permita en el futuro agregar nuevas columnas de datos a los existentes.
- En ambientes con información no estructurada, como por ejemplo las redes sociales, o donde la estructura de la información cambia regularmente, es más adecuado un formato JSON, donde la metadata está incluida en cada registro, esto permitiría agregar o remover columnas sin afectar la información existente.

En resumen de lo antes expuesto, podemos concluir que los portales web de datos abiertos del ámbito público deberían considerar el uso de modelos que permitan caracterizar y normalizar datos, como así también permita sugerir formatos de presentación junto con la exposición de una metadata asociada en pos de que los ciudadanos puedan hacer un buen uso de los datos públicos y a través de sus iniciativas innovadoras puedan aportar información de valor que permita mejorar la calidad de vida de los habitantes en las ciudades.

4. FORMACIÓN DE RECURSOS HUMANOS

Esta línea de investigación se está llevando a cabo en forma conjunta entre la Universidad Nacional de San Luis (a través de la Facultad de Ciencias Físico-Matemáticas y Naturales) y la Universidad Tecnológica Nacional (a través de sus Facultades Regionales de Córdoba y San Francisco).
El equipo de trabajo está compuesto por 3 docentes investigadores categorizados a nivel nacional, 2 Tesistas de Posgrado y 2 becarios de grado que se encuentran cursando actualmente la carrera de Ingeniería en Sistemas de Información.

5. BIBLIOGRAFÍA

[1] TIC - Tecnologías de la Información - http://aprendeenlinea.udea.edu.co/lms/investigacion/mod/page/view.php?id=3118.
[2] Cortagerena Alicia, Freijedo Cluadio – Tecnologías de la Información y la Comunicación – Editorial Prentice-Hall
[3] Maximiliano Bron - Universidad Nacional de La Rioja - 2015 - Open Data, Miradas y perspectivas de los datos abiertos - Obra colectiva y colaborativa. 1era Edición. ISBN: 978-987-1999-09-5.
[4] Calvin M.L. Chan - SIM University - From Open Data to Open Innovation Strategies: Creating e-Services Using Open Government Data. 2013 46th Hawaii International Conference on System Sciences.
[5] Ciudad Inteligente - http://www.endesaeduca.com/Endesa_educa/recursos-interactivos/smart-city
[6] Ciudad Inteligente - http://cintel.org.co/innovacion/ciudades-inteligentes.
[7] - Lorena BĂTĂGAN - Academic Economics Studies - Bucarest - Rumania - Open Data for Smart Cities - lorena.batagan@ie.ase.ro.
[8] Cátedra de Big Data - Universidad Tecnológica Nacional - Facultad Regional Córdoba. Profesor Titular: Ing. Calixto Maldonado. JTP: Ing. Franco Mana y Ayudante: Ing. Pablo Vaca.
[9] Definición de ETL - http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/312584/procesos-etl-definici-n-caracter-sticas-beneficios-y-retos.
[10] Ramez Elmasri, Shamkant Navathe - Fundamentos de bases de datos - Quinta Edición - Pearson / Addison Wesley.

[11] Neal Leavitt - Will NoSQL Databases Live Up to Their Promise?. TECHNOLOGY NEWS.
[12] Karamjit Kaur - Rinkle Rani - Modeling and Querying Data in NoSQL Databases. Computer Sci. and Engg. Deptt. Thapar University.
[13] José A. Senso, Antonio de la Rosa Piñero. El concepto de metadato. Algo más que descripción de recursos electrónicos.
[14] Twitter para desarrolladores - https://dev.twitter.com.
[15] Twitter - Red Social – https://twitter.com.
[16] JSON Org. http://json.org/example.html.
[17] Patrick Widener, Karsten Schwan y Greg Eisenhauer - College of Computing Georgia Institute of Technology - Atlanta, Georgia 30332-0280 - Open Metadata Formats: Efficient XML-Based Communication for Heterogeneous Distributed Systems.
[18] Huamin Wang y Zhiwei Ye - International School of Software Wuhan University y School of Computer Science Hubei University of Technology - Wuhan, P.R. China - An ETL Services Framework Based on Metadata.
[19] Aragón Open Data – http://opendata.aragon.es.
[20] Portal de Datos – CABA - https://data.buenosaires.gob.ar/.