LA CALIDAD DE LOS METADATOS COMO UNA LIMITACIÓN DE COMUNICACIÓN EN EL SECTOR OFICIAL

IVONNE RODRÍGUEZ-FLORES
Escuela Superior Politécnica de Chimborazo, Facultad de Informática y Electrónica, Ecuador.
irodriguez@espoch.edu.ec

CLAUDIA JIMÉNEZ RAMÍREZ
Universidad Nacional de Colombia, Departamento Ciencias de la Computación y Decisión, Colombia
csjimene@unal.edu.co

RESUMEN

La red Internet ha permitido el acceso a nuevas fuentes de datos, tales como las estadísticas oficiales y los datos gubernamentales abiertos. Este artículo presenta algunos aspectos en la calidad de los datos y metadatos que afectan la comunicación en el sector oficial considerando que una organización cualquiera hoy asume varios papeles en la gestión de los datos: de usuario, productor o proveedor. Mediante una revisión de literatura y analizando sitios Web de datos abiertos de Colombia, Ecuador, México y Perú, se plantea la necesidad de una gestión integral de la calidad de los datos abiertos y sus metadatos.

Palabras clave: calidad de datos, metadatos, datos abiertos, estadísticas oficiales.

1. INTRODUCCIÓN

La importancia de los datos, y el conocimiento que éstos pueden proporcionar, es cada vez más reconocido por las organizaciones (Baldwin, 2015; Experian, 2015). Esto se evidencia en el incremento de la demanda de analistas y personal para la gestión y análisis de los datos, hoy llamados *científicos de datos* (McKinsey & Company, 2011; Chen, Chiang, & Storey, 2012). Pero no sólo se requiere gestionar los datos generados por la organización, sino que también aquellos datos externos que, mediante procesos de análisis, contribuyan a la toma de decisiones basada en el conocimiento descubierto y así aplicar el enfoque gerencial conocido como Inteligencia de Negocios (Business Intelligence, en inglés y abreviadamente BI).

Los datos externos generalmente están dispersos en el espacio virtual, constituyéndose la red Internet como fuente muy importante de los datos. Entonces, podemos decir que el mundo virtual ahora es parte del mundo real de la organización, sin distinción del tamaño o al área de dominio a la que ésta corresponda (Rasmussen, 2008).

De acuerdo con Sundgren (2012) y el Grupo Asesor de Expertos Independientes (en sus siglas en inglés IEAG, 2014)[1], cualquier tipo de organización, asume tres papeles en la gestión y análisis de los datos: usuario, productor y proveedor. Por tal razón, pueden presentarse cambios en varios niveles organizacionales como en su operatividad, en el uso de las tecnologías informáticas, hasta en el nivel de formación de las personas. Es innegable que las organizaciones ahora se ven afectadas por los datos externos, ya que tradicionalmente se han mantenido en un contexto específico y cerrado.

Este artículo se centra en las entidades u organizaciones del sector oficial, y tiene por objetivo presentar cómo la falta de metadatos se convierte en una limitación de comunicación, considerando que ahora las entidades asumen nuevos papeles para dar cumplimiento a las regulaciones y políticas de gobierno abierto. En el presente trabajo, se consideran varias dimensiones de la calidad de datos y metadatos para evaluar su integralidad y su impacto en las comunicaciones. Dichas dimensiones son propuestas por dos comunidades de investigación: las oficinas de estadísticas nacionales y las organizaciones que proveen datos abiertos.

El artículo está organizado en cinco secciones. La primera sección se ha introducido el tema y presentado el objetivo del presente trabajo de investigación. En la segunda sección se presenta cómo la red Internet ha permitido el uso de nuevas fuentes de datos y por qué las estadísticas oficiales y los datos abiertos son importantes para las entidades gubernamentales. En la tercera sección se expone las limitaciones de comunicación, mientras que en la cuarta sección se presenta varias propuestas para evaluar la calidad de los datos y sus metadatos, desde un enfoque integral de gestión de la calidad de datos. Y finalmente, en la quinta sección, presentamos las conclusiones de este trabajo de investigación.

2. NUEVAS FUENTES DE DATOS

La ciencia de los datos y las iniciativas de gobierno abierto han incentivado a las organizaciones a utilizar datos externos que provienen de diversas fuentes, pero cuyo principal medio para obtenerlos es la red Internet. El uso de nuevas fuentes de datos, es una realidad que incluso las Naciones Unidas establece como uno de los "*Principios Fundamentales de las Estadísticas Oficiales*" la utilización de "*todo tipo de fuentes*", siempre que éstas se ajusten a criterios de calidad, oportunidad y costo (Principio 5) (United Nations, 2014).

La red Internet ha abierto nuevas vías para la recolección de datos que también requiere de nuevas formas de organización y difusión del conocimiento (Cooper, 2014). Por lo tanto, puede ser considerada como una fuente primordial de datos (Barcaroli et al., 2015; Berękesewicz, 2015), siendo a su vez, parte de la vasta categoría de grandes volúmenes de datos (Big Data, en inglés) según la clasificación propuesta por la Comisión Económica de las Naciones Unidas para Europa (UNECE-HLG-MOS, 2013) y las Naciones Unidas (UN-GWG, 2015).

El uso de internet para la recolección de datos lo podemos mirar desde dos puntos de vista: el primero es para la recolección de datos con propósito de producción (para generar parte de los datos internos), donde la red es un canal o medio de comunicación que sustituye a los instrumentos tradicionales como el teléfono, el cara a cara, entre otros. Por ejemplo, las encuestas

[1] Grupo Asesor de Expertos Independientes del Secretario General de las Naciones Unidas sobre la revolución de los datos para el desarrollo sostenible (IEAG).

web o encuestas por correo electrónico, son medios muy comunes de adquisición de datos (Rasmussen, 2008). El segundo punto de vista, se usa a la red Internet como nuevas fuentes de datos para la recolección de datos con el propósito de añadir datos para que cualquier tipo de organización pueda enriquecer los suyos (los internos) con datos externos de sitios web, de catálogos de datos abiertos, de estadísticas oficiales, y todos aquellos datos digitales que se puedan acceder y/o adquirir vía electrónica (correos electrónicos, archivos de bitácora o *logs* de transacciones, redes sociales, entre otros).

En el presente documento se consideran algunas fuentes de datos externas que provienen o se difunden en el sector gubernamental, como las estadísticas oficiales y los datos gubernamentales abiertos (DGA). Las dos fuentes son formas de datos abiertos, por lo tanto, su propósito es ser centros de acopio y publicación de conjuntos de datos socioeconómicos, que sean accesibles en formatos legibles por humanos o por máquinas, y que puedan ser reutilizados por cualquiera, libres de restricciones legales.

2.1 Estadísticas Oficiales y Datos Abiertos

La gran cantidad de datos y su demanda creciente han hecho que, desde la visión de las entidades internacionales tales como las Naciones Unidas, El Banco Mundial, la Comisión Económica de las Naciones Unidas para Europa (UNECE) y la Organización para la Cooperación y el Desarrollo Económicos (en sus siglas en inglés OCDE), se piense en utilizar a las Oficinas de Estadísticas Nacionales (OEN) como protagonistas de la "*Revolución de los datos*"[2] en los países en desarrollo (PARIS21, 2015). Es así que, varios proyectos que se iniciaron en torno a la "*Revolución de datos*", han generado grupos de trabajo enfocados a investigar los beneficios y retos de los grandes volúmenes de datos (Big Data) y su potencial para monitorear e informar sobre los objetivos de desarrollo sostenible como parte de la "*Agenda2030"* de las Naciones Unidas. Entre dichos grupos, se pueden mencionar al grupo para modernización de las estadísticas oficiales de la UNECE (UNECE-HLG-MOS)[3], y al grupo de trabajo global de las Naciones Unidas (UN-GWG)[4]. Estos grupos de trabajo se limitan a las Oficinas de Estadísticas Nacionales, pero se debe considerar que los datos también se producen y provienen de diferentes entidades, sectores o comunidades.

La revolución en datos en un país en desarrollo (que debe alinearse a Agendas internacionales, como la Agenda2030), considerando que las estadísticas oficiales serán las fuentes de datos claves para la toma de decisiones nacionales y el monitoreo internacional, hace que sea inevitable que las entidades, especialmente, las del sector oficial tengan que adaptarse al acceso y uso de nuevas fuentes de datos en un diferente ecosistema (IEAG, 2014). Una entidad, a más de ser productora de sus propios datos, se convierte en usuaria y proveedora de información que debe ser difundida como datos abiertos. De esta forma, contribuyendo con las estrategias y políticas de gobierno abierto.

Con la política de gobierno abierto, se pretende el involucramiento de la sociedad civil y los ciudadanos, de manera que puedan impulsar un desarrollo sostenible (IEAG, 2014). Las

[2] "Revolución de datos" es el enfoque en la que se orienta la Agenda2030 para el Desarrollo Sostenible que fue adoptada por los países miembros de las Naciones Unidas - Nueva York, 25 de septiembre, 2015.

[3] Big Data in Official Statistics, UNECE-HLG-MOS, http://www1.unece.org/stat/platform/display/bigdata/2015+project

[4] Big Data for Official Statistics, UN-GWG, https://unstats.un.org/bigdata/

estadísticas oficiales (producidas en una OEN) pueden presentarse a manera de datos abiertos en portales web, catálogos de datos abiertos u otro medio cualquiera. Pero la creciente sociedad digital hace que, incluso estas fuentes de datos estadísticos resulten insuficientes y no satisfagan la necesidad de información. Por lo que, existen cada vez más brechas en la cobertura estadística de los diferentes sectores económico y social, que son sólo parcialmente explicados por fuentes de datos estadísticos oficiales (Berękesewicz, 2015). Y surge la necesidad de recurrir a otras fuentes, como los catálogos de "*Datos Gubernamentales Abiertos*" (DGA).

Cabe señalar que los datos abiertos que las Oficinas de Estadísticas Nacionales (OEN) difunden, deben ser acompañadas por sus metadatos en dos niveles de abstracción: microdatos y macrodatos. Según el glosario de la Organización para la Cooperación y el Desarrollo Económicos, los microdatos son datos sobre las características de las unidades de una población, tales como individuos, hogares o establecimientos, recogidos por un censo, una encuesta o un experimento. Y los macrodatos, pueden definirse como datos derivados de los microdatos mediante estadísticas de grupos o agregados, como recuentos, medias o frecuencias (OECD-Glossary, 2016).

2.2 Los Metadatos

La National Information Standards Organization define a los metadatos como: "*información estructurada que describe, explica, localiza, o de otra manera lo hace más fácil para recuperar, utilizar o administrar un recurso de información*" (NISO, 2004). Sin embargo, la definición más conocida sobre metadatos en cualquier ámbito es: "*Datos acerca de los datos*", haciéndola más general e independiente del contexto, pero más ambigua.

En la actualidad la definición de metadatos también se ha popularizado para la organización de los recursos web con el propósito de facilitar la interoperabilidad y la integración de recursos, con el objeto de compartir o preservar la información. Por esto, en el ámbito web, Berners-Lee y el World Wide Web Consortium (W3C) nos presentan una definición más acotada y restrictiva de metadatos: "*es información comprensible por la máquina sobre recursos web u otras cosas*" (Berners-Lee, 1997).

Los metadatos son datos, por lo tanto se generan, se utilizan, se transforman o reutilizan de manera natural (Sundgren, 2012). Hoy en día, los metadatos no se restringen a un estado pasivo (documentación de objetos), sino también son utilizados en forma más activa lo que facilita la comunicación entre los usuarios (entre humanos o máquinas) y las fuentes de los datos (Lundell, 2013). Concisamente, los metadatos mantienen el registro sobre el significado, contexto y estructura de los datos, facilitando que éstos sean gestionados y analizados.

Se han definido estándares internacionales de metadatos con el fin de contar con información documentada en forma armonizada y de aceptación en el ámbito mundial. Sin embargo, los estándares pueden orientarse a dominios o aplicaciones específicos; tales como el "Estándar de Intercambio de Datos y Metadatos" (en sus siglas en inglés SDMX) para la comunidad Estadística. Otros estándares pueden usarse en varias comunidades, y en forma conjunta, pero con diferentes propósitos, tales como la "*Iniciativa de Documentación de Datos* (DDI)" para describir datos de las ciencias sociales, conductuales y económicas, y "*Dublin Core*" para la descripción y catalogación de objetos digitales en la Web.

Tanto las comunidades de las estadísticas como la de los datos abiertos, difunden los datos mediante portales oficiales o catálogos (software de terceros) e idealmente, estos datos deben estar acompañados de sus metadatos. Por lo tanto, un conjunto de datos no necesariamente se reduce únicamente a un solo archivo digital de datos, sino que necesita de otros complementarios.

3. LIMITACIONES DE COMUNICACIÓN

La organización de hoy en día afronta desafíos en torno a la *comunicación* entre humanos y/o máquinas. Una forma sencilla de comprender este punto es como lo plantea Bo Sundgren (2012) en su trabajo "*Communicating in time and space - How to overcome incompatible frames of reference of producers and users of archival data*". Para que haya comunicación debe haber un remitente (productor de datos – fuente) y un receptor que pueden ser bien diferentes y por lo tanto, se deben compensar las diferencias o incompatibilidades para la comprensión del mensaje o los datos que se están recibiendo. En el ámbito tecnológico, se deben considerar la variedad de datos, lenguajes, marcos, estándares, entre muchos otros factores que pueden incidir en la buena comunicación. Por esto, para facilitar que el receptor interprete adecuadamente el mensaje comunicado (datos e información), se requieren los metadatos, como se dijo antes, datos que describan o expliquen el significado de los datos.

El uso de datos externos, generalmente abiertos, hace que una entidad u organización enfrente desafíos de comunicación en un ecosistema diverso. Esto aplica por supuesto, en el sector oficial. Por ello, se exploraron algunos portales de las oficinas de estadísticas nacionales y los catálogos oficiales de datos abiertos (portales creados por los gobiernos para promover la producción y la utilización de datos abiertos) de cuatro países: Colombia, Ecuador, México y Perú (ver Tabla 1).

De la revisión a los portales oficiales de las oficinas nacionales de estadísticas y a los catálogos de datos gubernamentales abiertos que se presentan en la Tabla 1, se pudo observar lo siguiente:

1. Los cuatro países Colombia, Ecuador, México y Perú a través de sus Oficinas de Estadísticas Nacionales (abreviadamente, OEN) se han comprometido con agendas y metas internacionales relacionadas con la "*Revolución de los Datos*". Por esta razón, se pueden ver ciertas similitudes pues el Sistema Estadístico Nacional (SEN) de los cuatro países está conformado principalmente por dos plataformas, una de ellas propia de la OEN que se dedica a la difusión de las estadísticas oficiales (presentadas como mapas, tablas, gráficas, entre otras). Y la otra plataforma (provista por terceros), conocida como ANDA (abreviatura de Archivo Nacional de Datos) como herramienta de difusión y presentación de conjunto de datos (metadatos y microdatos).

   ANDA[5] es una herramienta de software libre desarrollada por la Red Internacional de Encuestas a Hogares con el apoyo del Banco Mundial, su dominio de aplicación es específico para las OENs y adopta estándares internacionales de metadatos para documentar los microdatos. Mediante el catalogo en línea ANDA, la OEN difunde el conjunto de datos que se conforma fundamentalmente de, un archivo con los microdatos (en formato: CSV, DBF, SAV, entre otros) y dos archivos de metadatos. El archivo de metadatos acorde al estándar "*Iniciativa de Documentación de Datos* (DDI)", describe las características importantes de los microdatos. Mientras que, el otro archivo con elementos del estándar "*Dublin Core*", describe

[5] http://www.ihsn.org/software/nada

los recursos relacionados a los microdatos, y su formato está en "*Marco de descripción de recursos* (RDF)". Adicionalmente, tanto DDI como Dublin Core usan el Lenguaje de Marcado eXtensible (XML).

En el portal web para estadísticas oficiales, se hace la difusión de macrodatos (a modo de indicadores) con formatos de descarga XLS y PDF como los más comunes; pero estos conjuntos de datos no contienen metadatos que ayuden a la comprensión de los datos por parte de los usuarios. También, estos portales comparten microdatos y el formato común utilizado corresponde al software estadístico comercial SPSS (formato SAV), y tampoco ofrecen metadatos adicionales de apoyo. Asimismo, los conjuntos de datos se duplican entre las dos plataformas (ANDA y Portal de Estadísticas Oficiales) en una misma OEN.

Tabla 1. Portales de oficinas de estadísticas nacionales y catálogos de datos gubernamentales abiertos

País	**Oficinas de Estadísticas Nacionales (OEN) y Sistema Estadístico Nacional (SEN)**		**Catálogos de Datos Gubernamentales Abiertos (DGA)**
	Estadísticas oficiales	**Catálogo de Datos abiertos**	
Colombia	*Plataforma*: Portal Web[6] *Formatos*: XLS, PDF	*Plataforma*: ANDA *Estándares de metadatos*: DDI, Dublin Core en RDF *Microdatos*: acceso restringido	*Plataforma:* SOCRATA[7] *Formatos*: CSV, XLS, JSON, XML, RDF
Ecuador	*Plataforma*: Portal Web[8] *Formatos*: Microdatos: SAV, CSV Macrodatos: XLS, PDF	*Plataforma:* ANDA *Estándares de metadatos:* DDI, Dublin Core en RDF *Microdatos*: SAV, CSV	*Plataforma:* CKAN[9] *Formatos*: CSV, XLS, JSON
México	*Plataforma:* Portal Web[10] *Formatos*: Microdatos: DBF, CSV, DTA, SAV, SAS Macrodatos: XLS, PDF	*Plataforma:* ANDA *Estándares de metadatos:* DDI, Dublin Core en RDF *Microdatos*: DBF, CSV, DTA, SAV, SAS	*Plataforma:* CKAN[11] *Formatos*: CSV, JSON
Perú	*Plataforma*: Portal Web[12] *Formatos*: Microdatos: SAV, DBF Macrodatos: XLS, PDF	*Plataforma:* ANDA *Estándares de metadatos:* DDI, Dublin Core en RDF *Microdatos*: SAV, DBF	*Plataforma:* Portal Web[13] *Estándares de metadatos:* Dublin Core. *Formatos:* CSV, JSON, RDF

Fuente: Elaboración propia.

[6] http://www.dane.gov.co/index.php/sistema-estadistico-nacional-sen

[7] https://www.datos.gov.co/

[8] http://www.ecuadorencifras.gob.ec/institucional/home/

[9] http://catalogo.datosabiertos.gob.ec/

[10] http://www.inegi.org.mx/

[11] https://datos.gob.mx/

[12] https://www.inei.gob.pe/sistema-estadistico-nacional/

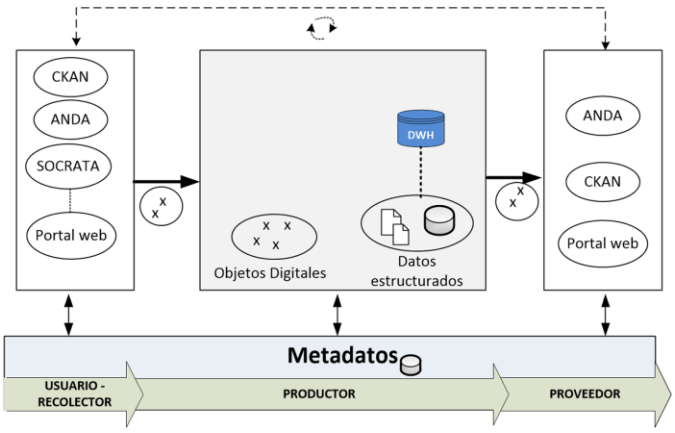
[13] http://www.datosabiertos.gob.pe/

2. Las organizaciones del sector público son las principales concentradoras y generadoras de información, impulsadoras de la política de "*Gobierno Abierto*" que ha asumido el sector oficial mundial y que en los últimos años se ha incrementado notablemente (CEPAL-UN, 2014). Se pone a disposición del público en general como "datos gubernamentales abiertos". Y se divulgan a través de catálogos que pueden ser sus propios portales Web, como en el caso de Perú, o mediante sitios de terceros como CKAN[14] (de comunidad) para Ecuador y México; mientras que Colombia utiliza SOCRATA (comercial).

   CKAN es una plataforma de código abierto y libre desarrollada por la Fundación de Conocimiento Abierto (de sigla en inglés, OKF). Es la más utilizada en el sector gubernamental (Oliveira, de Oliveira, Oliveira, & Lóscio, 2016). Los conjuntos de datos del CKAN incluyen metadatos generales para describir información contextual importante sobre los datos y sus fuentes. Los metadatos de los conjuntos de datos se publican de forma nativa en formato "*Notación de Objetos de JavaScript* (JSON)". Sin embargo, la plataforma también permite otros formatos comunes como CSV (archivos separados por comas) y de hojas electrónicas.

La Figura 1 muestra gráficamente el escenario en que se mueven las entidades gubernamentales en los entornos de datos abiertos examinados. En resumen, la entidad inicia como *usuario*, y la recolección se realiza desde portales oficiales de datos abiertos; en primera instancia se acude al portal web del SEN y al catálogo ANDA, ambas herramientas son de la Oficina de Estadísticas Nacionales. Seguidamente, se puede recurrir al catálogo oficial de datos abiertos; el cual, usualmente es el software CKAN (principalmente en los países de América del Sur). En este escenario, resulta un desafío la diversidad de estándares y formatos, que además hacen que los lenguajes para intercambio de datos también varíen. Así tenemos que, para los metadatos de las OENs utilizan XML (ANDA), mientras que los catálogos de datos gubernamentales abiertos utilizan JSON (CKAN).

Figura 1. Ecosistema para el uso de nuevas fuentes de datos



Fuente: Elaboración propia.

[14] https://ckan.org/

La entidad en su papel de *productor* deberá almacenar y analizar los datos recopilados, e integrarlos con sus propios datos (internos). En su tercer papel, la entidad se convierte en *proveedor*, donde es responsable de la difusión de datos e información a través de su portal oficial, o un catálogo como el CKAN o ANDA; y de esta manera se ha formado un circuito. Las etapas que abarcan los tres papeles de la organización (usuario, productor o proveedor), deben ser cubiertas transversalmente con la gestión de los metadatos; sin olvidar que éstos, fundamentalmente deben ser de alta calidad.

4. CALIDAD DE LOS DATOS Y METADATOS

La definición general de calidad de datos es "*apropiados para el uso*" (Wang & Strong, 1996; Redman, 2013). Por otro lado, está definición por sí sola, no presenta alguna orientación sobre cómo medir lo "*apropiado*" o cómo decidir si algunos datos son inadecuados (no apropiados). Entonces, puede resultar en una definición subjetiva incluso relativa al momento en que se tenga que evaluar la calidad de los datos (Rasmussen, 2008). La norma internacional ISO/IEC 25012:2008 presenta otra definición de calidad de datos, como el: "*grado en que las características de los datos satisfacen las necesidades expresadas e implícitas cuando se utilizan en condiciones especificadas*" (ISO/IEC 25012, 2008).

El ejemplo descrito de la sección anterior muestra los desafíos que desde la perspectiva de una organización debe enfrentar. Y el desafío de interés para este artículo, es la calidad de los datos, tanto de los externos que provienen de las nuevas fuentes de datos, como de los generados por ella misma. No obstante, sin distinción de dominios de aplicación, los datos y sus metadatos deben ser de calidad. Y, la calidad de datos debe gestionarse para garantizar su confiabilidad.

La calidad de datos ha sido objeto de investigación por varias décadas (Scannapieco & Catarci, 2002; Sadiq, Yeganeh, & Indulska, 2011; Illari & Floridi, 2014; Batini & Scannapieco, 2016). Sin embargo, recientes estudios realizados por las comunidades de las Oficinas de Estadísticas Nacionales (Daas & Ossen, 2011; ESSnet-Data warehouse, 2013; IEAG, 2014), y de Datos Abiertos (Zuiderwijk et al., 2012; Reiche, Hofig, & Schieferdecker, 2014; Umbrich, Neumaier, & Polleres, 2015; Oliveira et al., 2016) aún reportan problemas de calidad de los datos y sus metadatos. Y éste hecho, es consecuencia de tratar con grandes volúmenes, una alta variedad de datos, con islas de datos (internas o externas) y el uso de diversas tecnologías y estándares. Tal como se describió en la sección anterior.

La calidad de los datos se mide a través de múltiples dimensiones (o características en ISO/IEC 25012). Una dimensión es una propiedad medible de calidad del dato que representa algún aspecto del dato que se pueden utilizar para guiar el proceso de comprensión de la calidad (Wang & Strong, 1996; Batini & Scannapieco, 2006). Puede ser que ciertos datos particulares puedan ser descritos como de alta calidad, de acuerdo con una o más dimensiones. Sin embargo, la calidad de datos por ser de carácter multidisciplinario y relativo al contexto, hace que se encuentren diferentes términos que se refieren a la misma dimensión o viceversa.

Con base en lo expuesto, para una gestión integral de la calidad de datos y metadatos, se debe lograr una adecuada sinergia entre propuestas de diferentes comunidades de investigación en este campo (Sadiq et al., 2011). Por tal razón, se han seleccionado varias propuestas de las Oficinas

de Estadísticas Nacionales y de Datos Abiertos (ver Tablas 2 y 3) que se ajustan al escenario real de interés, además se establece la aplicabilidad en una organización ya sea como: usuaria, productora o proveedora de datos.

Tabla 2. Dimensiones de calidad de Metadatos: Comunidad de Oficinas de Estadísticas Nacionales

Autores	Categorías	Dimensiones	Perspectiva
Sistema Estadístico Europeo (ESS)	x	Relevancia Exactitud Oportunidad y Puntualidad Accesibilidad y Claridad Comparabilidad Coherencia	Producción y difusión de estadísticas (*productor y proveedor*)
Daas & Ossen (2011)	Fuente	Proveedor Relevancia Privacidad y Seguridad Entrega Procedimientos	Fuentes de datos secundarios o "administrativos" (*usuario*)
	Metadatos	Claridad Comparable Claves únicas Tratamiento de datos por parte del responsable de la fuente de datos	
	Datos	Verificación Técnica Exactitud Completitud Dimensión relacionada a Tiempo Integrable	
Proyecto: Enlazar microdatos y el sistema de almacenamiento de datos en la producción estadística[15] (ESSnet-Data warehouse, 2013)	Fuente	Relevancia Completitud Exactitud	Gestión de metadatos (microdatos) con propósito de Producción y difusión de estadísticas. *(usuario, productor y proveedor*)
	Integración	Coherencia Unicidad (único)	
	Interpretación y Análisis	Relevancia Entendimiento	
	Acceso	Correcto (Exactitud)	

Fuente: Elaboración propia.

Las dimensiones que se muestran en la Tabla 2, son propuestas por investigadores u organismos internacionales de las oficinas de estadísticas nacionales. De estas propuestas se resaltan los siguientes aspectos:

i) Las dimensiones del Sistema Estadístico Europeo (siglas en inglés ESS) son la base para muchas OEN tanto de Europa como de los países miembros de la UNECE. Por otro lado, estas dimensiones han sido propuestas desde una perspectiva de producción y difusión de estadísticas.

ii) En el trabajo de Daas & Ossen (2011), se propone una evaluación de la calidad desde una perspectiva de fuentes de datos secundarios o "administrativos" (nuevas fuentes – externas),

[15] http://ec.europa.eu/eurostat/cros/content/data-warehouse_en

y las dimensiones las agrupa en tres hiperdimensiones que son: fuente, metadatos y datos. A las dimensiones se asocian indicadores de calidad, cuyo cumplimiento se realiza mediante una lista de verificación.

Esta propuesta está centrada a controlar la calidad de las nuevas fuentes de datos (llamadas secundarias por el autor), por lo que desde una perspectiva de la organización, se enfocaría únicamente en su papel de *usuario*, y necesitaría ser complementada.

iii) Como Resultado del proyecto "*Enlazar microdatos y el sistema de almacenamiento de datos en la producción estadística*", se tiene un marco para la gestión de metadatos (para microdatos) mediante una Bodega de datos (Data Warehouse). Las dimensiones propuestas se agrupan según las capas que están distribuidas en la arquitectura de almacenamiento de los datos estadísticos. Y estas capas son: Fuente, Integración, Interpretación y análisis, y la de Acceso. El dominio de aplicación del marco es específico al estadístico, ya que se sustenta en reglas y estándares como: el "Estándar de Intercambio de Datos y Metadatos" (SDMX, por su sigla en inglés) y el modelo de metadatos nórdico, entre los principales.

Esa última propuesta tiene una mayor cobertura de gestión y análisis, en relación con las anteriores, de manera que se puede considerar que la organización puede cumplir con los tres papeles ya mencionados.

Las dimensiones de la Tabla 3, corresponden a propuestas de investigadores de la comunidad de datos abiertos. A continuación, se realiza una descripción de dichas propuestas.

i) Los trabajos de Umbrich, Neumaier & Polleres (2015, 2016), proponen un marco para evaluación y monitoreo de la calidad de los portales de datos abiertos en forma automática. La recuperación de los metadatos se limita a software de terceros (catálogos), que son: CKAN, Socrata, OpenDataSoft (Neumaier, Umbrich, & Polleres, 2016).

Se considera a esa última propuesta que, desde la perspectiva de la organización, puede actuar como *usuaria* porque recupera los metadatos y los almacena en una base de datos en Postgres, y con algunos módulos desarrollados en Python, evalúa con métricas, la calidad de los metadatos considerando sus múltiples dimensiones. Y a través de un cuadro de mando, reporta de forma automática los resultados de la evaluación. También se le ha considerado como *proveedor* porque mediante una interfaz de usuario, la herramienta tecnológica[16], pone a disposición los metadatos en su estado natural (crudo) para su reutilización.

ii) Las dimensiones propuestas por Reiche, Höfig, & Schieferdecker (2014) se derivan de la evaluación automática a catálogos CKAN de datos gubernamentales abiertos (DGA), y los resultados de dicha evaluación son agregados y visualizados a través de una aplicación web, con el propósito de proveer un servicio de monitoreo continuo. Por esta razón, se lo ubica con una perspectiva de organización como *Usuario*. Sin embargo, el mayor aporte de los autores es poder cuantificar la calidad de los metadatos de los portales de datos abiertos, mediante métricas formuladas.

[16] http://data.wu.ac.at/portalwatch/

Tabla 3. Dimensiones de calidad de Metadatos: Comunidad de Datos Abiertos

Autores	Categorías	Dimensiones	Perspectiva
Umbrich, Neumaier & Polleres (2015, 2016)	Metadatos	Recuperable Uso Completitud Exactitud Apertura Localizable	Evaluar portales de datos abiertos (*usuario*, *proveedor*)
Reiche, Höfig, & Schieferdecker (2014)	Metadatos	Completitud Completitud ponderada Exactitud Riqueza de información Legibilidad Disponibilidad Error ortográfico	Evaluar portales de datos abiertos (*usuario*)

Fuente: Elaboración propia.

5. CONCLUSIONES

El uso de nuevas fuentes de datos hace que las organizaciones enfrenten limitaciones de comunicación. Esto es previsible ya que tradicionalmente se ha mantenido en un contexto específico y cerrado que constituye la realidad que está en la mente de sus funcionarios. La calidad de los datos y metadatos es fundamental, ya que, desde una perspectiva de la organización como *usuaria*, debe confiar en los datos externos que está recolectando e integrando a los suyos, para que luego puedan ser analizados e interpretados como *productora*. Y por último, garantizar que los datos que se producen en la organización también sean correctos y puedan ser difundidos como datos abiertos en su papel de *proveedora de datos*.

La diversidad de esquemas, estándares y tecnologías, fuerza a nuestras instituciones gubernamentales a plantearse la gestión integral de datos y metadatos como una necesidad. Para lo cual, es necesario que las buenas prácticas de ambas comunidades, las oficinas de estadísticas nacionales y los datos abiertos; se relacionen y adapten en torno a los requerimientos propios (contexto) de la organización. Y justamente, éste es el trabajo futuro que se plantea para continuar con la investigación.

REFERENCIAS

Baldwin, H. (2015). Drilling Into The Value Of Data. Retrieved September 5, 2016, from http://www.forbes.com/sites/howardbaldwin/2015/03/23/drilling-into-the-value-of-data/print/

Barcaroli, G., Nurra, A., Salamone, S., Scannapieco, M., Scarnò, M., & Summa, D. (2015). Internet as Data Source in the Istat Survey on ICT in Enterprises. *Austrian Journal of Statistics*, *44*(2), 31–43. https://doi.org/10.17713/ajs.v44i2.53

Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques* (1st ed.). Secaucus, NJ, USA: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-33173-5

Batini, C., & Scannapieco, M. (2016). *Data and Information Quality - Dimensions, Principles and Techniques*. (M. J. Carey & S. Ceri, Eds.) (1st ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-24106-7

Berękesewicz, M. (2015). Estimating The Size Of The Secondary Real Estate Market Based On Internet Data Sources. *Folia Oeconomica Stetinensia*, *14*(2), 259–269. https://doi.org/10.1515/foli-2015-0012

Berners-Lee, T. (1997). Web architecture: Metadata. Retrieved July 15, 2016, from https://www.w3.org/DesignIssues/Metadata.html

CEPAL-UN. (2014). *United Nations E-Government Surveys*. New York, NY, USA: United Nations.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188. Retrieved from http://dl.acm.org/citation.cfm?id=2481674.2481683

Cooper, P. (2014). Data, information, knowledge and wisdom. *Anaesthesia & Intensive Care Medicine*, *15*(1), 44–45. https://doi.org/10.1016/j.mpaic.2013.11.009

Daas, P. J. H., & Ossen, S. J. L. (2011). Metadata quality evaluation of secondary data sources. *Proceedings of 5th International Quality Conference*, 823–836. Retrieved from http://www.cqm.rs/2011/cd/5iqc/pdf/096.pdf

ESSnet-Data warehouse. (2013). *Metadata Framework for Statistical Data Warehousing*. Retrieved from http://ec.europa.eu/eurostat/cros/content/dwh-sga2-wp1-11-metadata-framework-statistical-data-warehousing-v112-final_en

Experian, D. Q. (2015). *The data quality benchmark report*. *Experian Data Quality*. Boston, MA, EEUU. Retrieved from http://cdn.qas.com/us-marketing/whitepapers/data-quality-benchmark-report-2015.pdf

IEAG. (2014). *A World that Counts. Mobilising the data revolution for sustainable development*. *UN report*. Retrieved from www.undatarevolution.org

Illari, P., & Floridi, L. (2014). Information Quality, Data and Philosophy. In L. Floridi & P. Illari (Eds.), *The Philosophy of Information Quality SE - 2* (Vol. 358, pp. 5–23). Springer International Publishing. https://doi.org/10.1007/978-3-319-07121-3_2

ISO/IEC 25012. (2008). Software engineering - Software product Quality Requirements and Evaluation (SQuaRE)- Data quality model. Retrieved September 1, 2016, from http://www.iso.org/iso/catalogue_detail.htm?csnumber=35736

Lundell, L.-G. (2013). *Framework of metadata requirements and roles in the Data warehouse Statistical*. *ESSnet on micro data linking and data warehousing in statistical production*. UNECE. Retrieved from http://www.cros-portal.eu/content/dwh-sga2-wp1-11-metadata-framework-statistical-data-warehousing-v112-final

McKinsey & Company. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, (June), 156. https://doi.org/10.1080/01443610903114527

Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated Quality Assessment of Metadata Across Open Data Portals. *J. Data and Information Quality*, *8*(1), 2:1--2:29. https://doi.org/10.1145/2964909

NISO. (2004). *Understanding Metadata*. (NISO Press, Ed.), *American Political Science Review* (1st ed.). MD, USA: National Information Standards Organization. Retrieved from http://www.mendeley.com/catalog/understanding-metadata-3/%5Cnhttp://www.niso.org/publications/press/UnderstandingMetadata.pdf%5Cnhttp://www.journals.cambridge.org/abstract_S0003055403000534

OECD-Glossary. (2016). The OECD Glossary of Statistical Terms. Retrieved May 23, 2016, from https://stats.oecd.org/glossary/

Oliveira, M. I. S., de Oliveira, H. R., Oliveira, L. A., & Lóscio, B. F. (2016). Open Government Data Portals Analysis: The Brazilian Case. In *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research* (pp. 415–424). New York, NY, USA: ACM. https://doi.org/10.1145/2912160.2912163

PARIS21. (2015). *A Road Map for a Country-led Data Revolution*. OECD Publishing. https://doi.org/10.1787/9789264234703-en

Rasmussen, K. B. (2008). General approaches to data quality and Internet-generated data. In F. Nigel G (Ed.), *The Sage handbook of online research methods* (pp. 79–96).

Redman, T. (2013). Data Quality Management Past, Present, and Future: Towards a Management System for Data. In S. Sadiq (Ed.), *Handbook of Data Quality SE - 2* (pp. 15–40). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36257-6_2

Reiche, K. J., Höfig, E., & Schieferdecker, I. (2014). Assessment and Visualization of Metadata Quality for Open Government Data. In *In International conference for e-democracy and open government.* (pp. 335–346). Austria. https://doi.org/978-3-902505-35-4

Sadiq, S., Yeganeh, N. K., & Indulska, M. (2011). 20 years of data quality research: themes, trends and synergies. *Proceedings of the Twenty-Second Australasian Database Conference - Volume 115*, 153–162. Retrieved from http://dl.acm.org/citation.cfm?id=2460396.2460415

Scannapieco, M., & Catarci, T. (2002). Data quality under a computer science perspective. *Archivi & Computer*, *2*,

1–15.
Sundgren, B. (2012). Communicating in Time and Space: How to Overcome Incompatible Frames of Reference of Producers and Users of Archival Data. *DDI Working Paper Series*, 1–25. https://doi.org/10.3886/DDIOtherTopics03
Umbrich, J., Neumaier, S., & Polleres, A. (2015). Quality assessment and evolution of Open Data portals. In *Proceedings - 2015 International Conference on Future Internet of Things and Cloud, FiCloud 2015 and 2015 International Conference on Open and Big Data, OBD 2015* (pp. 404–411). IEEE. https://doi.org/10.1109/FiCloud.2015.82
UN-GWG. (2015). *Revision and Further Development of the Classification of Big Data*.
UNECE-HLG-MOS. (2013). Classification of Types of Big Data. Retrieved March 27, 2016, from http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data
United Nations. Fundamental Principles of Official Statistics, Pub. L. No. A/RES/68/261, 28 (2014). UN General Assembly Resolution 68/261. https://doi.org/10.1093/oxfordhb/9780199560103.003.0005
Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33. Retrieved from http://www.jstor.org/stable/40398176?origin=JSTOR-pdf&seq=1#page_scan_tab_contents
Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., Alibaks, R. S., & Sheikh_Alibaks, R. (2012). Socio-technical impediments of open data. *Electronic Journal of E-Government*, *10*(2), 156–172.