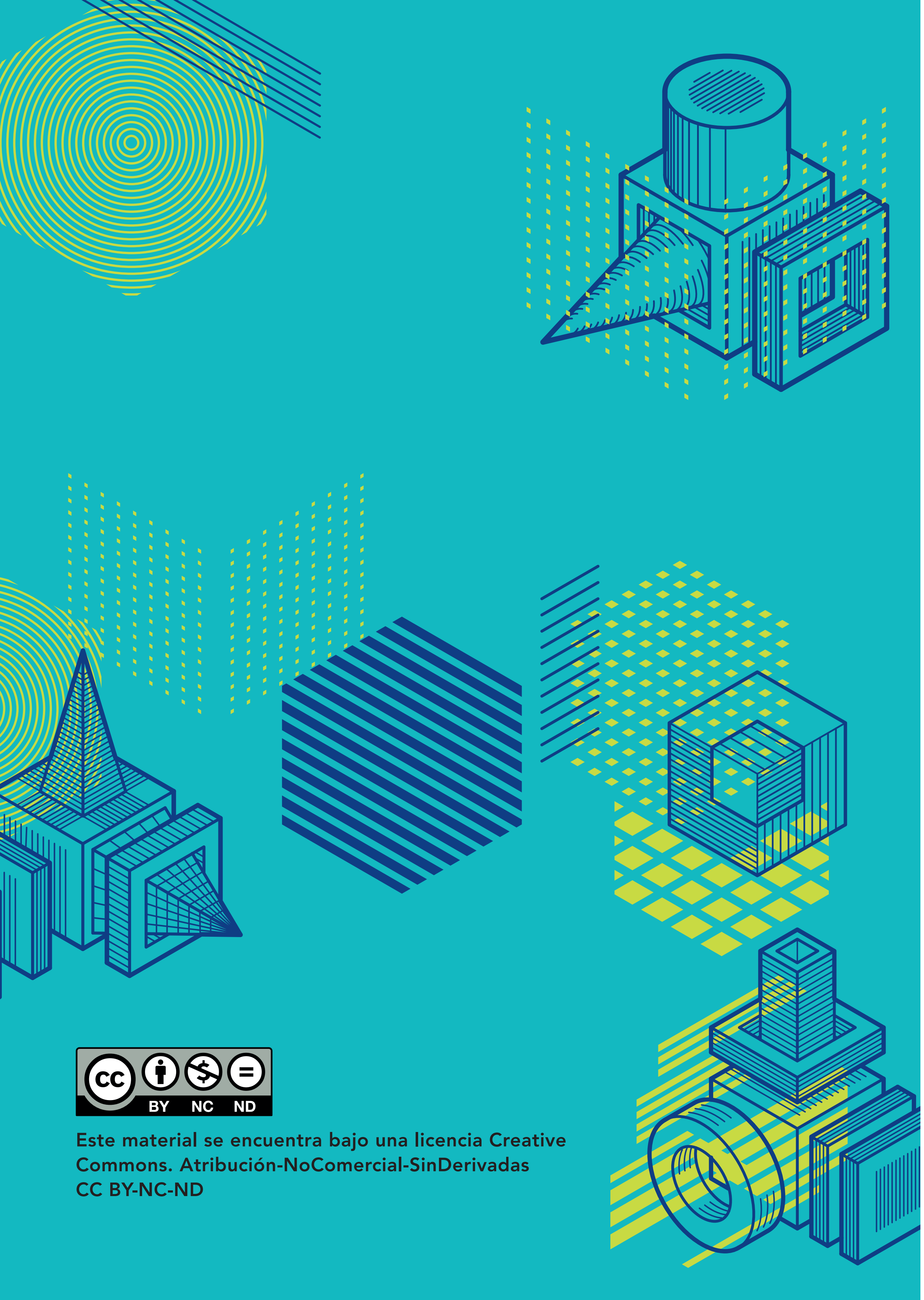
FUNDAMENTOS PARA LA PUBLICACIÓN DE DATOS EN LA WEB

ceweb.br nic.br cgi.br

TRUST
THE TRUST FOR THE AMERICAS

OEA
Más derechos para más gente



FUNDAMENTOS PARA LA PUBLICACIÓN DE DATOS EN LA WEB

Bernadette Farias Lóscio (UFPE)
Caroline Burle (Ceweb.br/NIC.br)
Marcelo Iury S. Oliveira (UFRPE)
Newton Calegari (Ceweb/NIC.br)

CGI.br
Comité Gestor de
Internet en Brasil
2018

Este material fue desarrollado por el **Centro de Estudios sobre Tecnologías Web del Núcleo de Información y Coordinación del Punto BR (Ceweb/NIC.br)** en el marco del proyecto "De un Gobierno Abierto a un Estado Abierto", ejecutado por **The Trust for the Americas** de la **Organización de los Estados Americanos (OEA)** con financiamiento de la Embajada de Estados Unidos en San José, Costa Rica.

ceweb.br nic.br cgi.br

NÚCLEO DE INFORMACIÓN Y COORDINACIÓN DEL PUNTO BR – NIC.BR

Director Presidente: Demi Getschko
Director Administrativo: Ricardo Narchi
Director de Servicios y Tecnología: Frederico Neves
Director de Proyectos Especiales y de Desarrollo: Milton Kaoru Kashiwakura
Director de Asesoría a las Actividades de CGI.br: Hartmut Richard Glaser

CENTRO DE ESTUDIOS SOBRE TECNOLOGÍAS EN LA WEB – CEWEB.BR

Organización: Beatriz Rossi Corrales
Equipo técnico: Amanda Marques, Beatriz Rossi Corrales, Caroline Burle, Diogo Cortiz, Mariana Frizanco, Newton Calegari, Reinaldo Ferraz y Selma de Morais
Revisión: Caroline Burle, Bernadette Farias Lóscio y Beatriz Rossi Corrales
Producción: Caroline D'Avo (Comunicación NIC.br) y Everton Rodrigues (Comunicación NIC.br)
Proyecto gráfico e ilustración: Giuliano Galvez (Comunicación NIC.br)

AUTORES

Bernadette Farias Lóscio
Centro de Informática – Universidad Federal de Pernambuco (UFPE)
bfl@cin.ufpe.br

Caroline Burle
Centro de Estudios sobre Tecnologías en la Web (Ceweb.br)
Núcleo de Coordinación de Información del Punto Br (NIC.br)
cburle@nic.br

Marcelo Iury S. Oliveira
Unidad Académica de Serra Talhada –Universidad Federal de Pernambuco (UFPE)
marcelo.iury@ufrpe.br

Newton Calegari
Centro de Estudios sobre Tecnologías en la Web (Ceweb.br)
Núcleo de Coordinación de Información del Punto Br (NIC.br)
newton@nic.br

ÍNDICE

INTRODUCCIÓN

Desde sus inicios, la Web se ha destacado como un importante medio para el intercambio de información. En este escenario con una gran cantidad de datos disponibles en la Web, hay dos papeles que merecen ser destacados: los publicadores y los consumidores de datos. En términos generales, los publicadores de datos tienen como objetivo la publicación y el intercambio de datos, con acceso libre o controlado, mientras que los consumidores de datos (que a la vez también pueden ser publicadores) desean utilizar estos datos para generar información útil y pertinente, así como para generar nuevos datos.

Es importante resaltar que el interés en publicar datos en la Web no es nuevo (BERNERS-LEE; CONNOLLY; SWICK, 1999 y ABITEBOUL; BUNEMAN; SUCIU, 2000). Sin embargo, en los últimos años, este interés se ha caracterizado por la publicación de datos de manera de promover su intercambio y reutilización. Así, no alcanza simplemente con poner a disposición el acceso a los datos. En general, los datos se deben publicar de forma que los consumidores los puedan comprender y utilizar fácilmente y que, además, estén disponibles en formatos que las aplicaciones puedan procesar con facilidad. Sin embargo, la heterogeneidad de los datos y la falta de estándares para la descripción y el acceso a los conjuntos de datos hacen que el proceso de publicar, compartir y consumir datos sea una tarea compleja. En este contexto, este artículo discute los fundamentos relacionados con la publicación de los datos en la Web, abordando aspectos relevantes, entre ellos los conceptos de Datos Abiertos, Datos Conectados (del inglés *Linked Data*), Ciclo de Vida de los Datos en la Web, y Buenas Prácticas para los Datos en la Web.

DATOS ABIERTOS

De acuerdo con la Open Knowledge Foundation (2012), Dato Abierto es cualquier dato que cualquier persona puede utilizar, reutilizar y redistribuir libremente. De este modo, los datos abiertos consisten en la publicación y diseminación de información en Internet, compartida en formatos abiertos, legible por máquinas y que pueda ser libremente reutilizada por la sociedad de forma automatizada. Es decir, a la apertura de los datos le interesa evitar un mecanismo de control y restricciones sobre los datos publicados, permitiendo que personas tanto físicas como jurídicas puedan explotar libremente dichos datos (ISOTANI; BITTENCOURT, 2015). Un dato se considera abierto cuando presenta las siguientes características (OPEN KNOWLEDGE, 2012):

I. Disponibilidad y acceso: el dato debe estar disponible en su totalidad. Debe estar en un formato conveniente y modificable;
II. Reutilización y redistribución: el dato debe ser suministrado en condiciones de reutilización y redistribución, y se debe poder combinar con otros;
III. Participación universal: todos pueden utilizar, reutilizar y redistribuir el dato sin restricciones de áreas, personas o grupos.

Los datos abiertos se pueden clasificar de acuerdo con una escala basada en estrellas propuesta por Tim Berners-Lee (BERNERS-LEE, 2006). Según esta clasificación presentada en la Figura 1, un dato publicado en la Web en cualquier formato (imagen, tabla o documento) y asociado a una licencia que permita su uso y reutilización sin restricciones se clasifica con *1 Estrella*. A pesar de que ya es un avance, los datos con *1 Estrella* deben ser manipulados en forma manual o por medio de extractores construidos específicamente para acceder a los datos.

Figura 1: Esta ilustración se basa en el esquema propuesto por Tim Berners-lee (2006)

A partir del momento en que los datos se publican en un formato que puede ser procesado automáticamente por algún software (por ejemplo, una hoja de cálculo Excel en vez de una imagen), los datos pasan a ser clasificados como *2 Estrellas*. Por un lado, esto puede facilitar el trabajo del consumidor de datos, aunque, por otro lado, puede dificultar un poco la tarea de publicación.

Los datos reciben la clasificación de *3 Estrellas* cuando se publican en formatos no propietarios (por ejemplo, CSV en vez de Excel). Nuevamente, la publicación de datos en formatos abiertos puede implicar costos adicionales para los publicadores. Esto ocurre cuando el formato de origen es diferente al formato adoptado para la publicación y es necesario convertir los datos y mantener la coherencia entre la fuente de datos original y los datos publicados en formato abierto.

Cuando los datos reciben una identificación única y se conectan con otros datos, estos datos se pueden clasificar como *4 Estrellas*. La creación de vínculos entre los datos les permite formar parte de una red más amplia de datos abiertos y conectados (BIZER; HEATH; BERNERS-LEE, 2009). Por último, los datos reciben la clasificación de *5 Estrellas* si están conectados con otros datos ya disponibles en la Web. En este caso, es necesario identificar datos que representan el mismo concepto a fin de establecer los vínculos entre ellos.

Siguiendo el movimiento de los datos abiertos, los gobiernos de diferentes países están usando la Web como un medio para la publicación de datos e información sobre sus administraciones. Denominados Datos Abiertos Gubernamentales, estos datos se pueden encontrar fácilmente en los así llamados Portales de Datos Abiertos, los cuales ofrecen una interfaz más amigable para catalogar y acceder a los datos. Como ejemplos de portales de datos abiertos ya consolidados, se destacan el portal de EE.UU [1]. y el portal del Reino Unido [2]. Diversos países de Europa, como Francia [3] y Holanda [4], así como algunos países de América Latina, como Chile [5] y Uruguay [6], también tienen portales de datos gubernamentales abiertos. En el caso de

http://data.gov [1]
http://data.gov.uk [2]
http://data.gouv.fr [3]
http://dataoverheid.nl [4]
http://datos.gob.cl [5]
http://datos.gub.uy [6]

Brasil, el portal de datos abiertos [7] se lanzó a principios de 2012, siendo una iniciativa liderada por el Ministerio de Planificación.

La iniciativa de la apertura de los datos por parte de los gobiernos se ha visto impulsada por la búsqueda de transparencia, colaboración y participación de la sociedad y la comunidad (GOLDSTEIN; DYSON, 2013). Con el objetivo de alcanzar consenso sobre los requisitos necesarios para caracterizar una base de datos abiertos, el grupo de trabajo *Open Government Working Group* elaboró los ocho principios de los datos gubernamentales abiertos (TAUBERER; LESSIG, 2007):

- **Completos:** todos los datos deben estar disponibles sin limitaciones. Un dato público es aquel que no está sujeto a limitaciones válidas de privacidad, seguridad o privilegios de acceso.

- **Primarios:** los datos deben estar en formato bruto, sin agregados o modificaciones.

- **Actuales:** los datos deben ser publicados tan rápidamente como sea necesario para preservar su valor.

- **Accesibles:** los datos deben ser accesibles por el mayor número posible de usuarios y para el mayor número posible de propósitos.

- **Procesables por máquinas:** los datos deben estar razonablemente estructurados para permitir su procesamiento automatizado.

- **No discriminatorios:** los datos deben estar disponibles para todos, sin necesidad de registro.

[7] http://dados.gov.br

- **No propietarios:** los datos deben publicarse en un formato abierto sobre el cual ninguna entidad tenga control exclusivo.

- **Licencias libres:** los datos no deben estar sujetos a ninguna normativa sobre derechos de autor, patentes, propiedad intelectual o secreto industrial. Se pueden permitir restricciones prudentes relacionadas con la privacidad, la seguridad y los privilegios de acceso.

Los datos gubernamentales abiertos se refieren a diversos temas y pueden implicar desde datos sobre los gastos y los ingresos del gobierno hasta datos sobre un censo escolar, puntos turísticos, reclamos de consumidores, demandas de servicios y muchos otros. En general, los datos disponibles provienen de actividades rutinarias realizadas dentro de los órganos gubernamentales, como los ministerios y las secretarías.

Una vez que los datos gubernamentales están disponibles en formato abierto, se espera que sean utilizados en el desarrollo de aplicaciones accesibles que puedan ser utilizadas fácilmente tanto por los ciudadanos como por el propio gobierno. Las aplicaciones ofrecen medios para analizar los datos por medio de filtros y también permiten visualizar los datos de forma simple y creativa. Ya hay diferentes aplicaciones y visualizaciones disponibles en la Web, resultado principalmente de concursos y *hackathons* promovidos para divulgar y popularizar los portales de datos abiertos.

DATOS CONECTADOS

El concepto de Datos Conectados se puede definir como un conjunto de Buenas Prácticas para publicar y conectar conjuntos de datos estructurados en la Web, con el fin de crear una "Web de datos" (BIZER; HEATH; BERNERS-LEE, 2009). La Web de datos genera innumerables oportunidades para la integración semántica de los propios datos, lo que promueve el desarrollo de nuevos tipos de aplicaciones y herramientas, como navegadores y motores de búsqueda (ISOTANI; BITTENCOURT, 2015).

Para comprender mejor la Web de datos, se puede establecer un paralelo entre la Web de Documentos (es decir, la Web actual) y la Web de Datos. La primera utiliza el estándar HTML para acceder a los datos, mientras que en la segunda a los datos se accede a partir del estándar RDF (ISOTANI; BITTENCOURT, 2015). La Web de documentos se basa en un conjunto de estándares que incluye: un mecanismo de identificación global y único, los URI (*Uniform Resource Identifiers*); un mecanismo de acceso universal, el HTTP; y un formato estándar para la representación de contenido, el HTML. De manera similar, la Web de Datos también se basa en una serie de estándares, entre ellos: el mismo mecanismo de identificación y acceso universal que se utiliza en la Web de Documentos (URIs y HTTP, respectivamente); un modelo estándar para la representación de los datos, el RDF; y un lenguaje de consulta para acceder a los datos, el lenguaje SPARQL (ISOTANI; BITTENCOURT, 2015).

Los principios de Datos Conectados fueron introducidos por Tim Berners-Lee (2006) y se resumen en cuatro principios básicos:

I. Utilizar URIs como nombre para los recursos;
II. Utilizar URIs HTTP para que las personas puedan encontrar estos nombres;
III. Cuando se accede a un URI, garantizar que se pueda obtener información útil de dicho URI, la cual debe estar representada en formato RDF;
IV. Incluir enlaces a otros URI de forma que se puedan descubrir otros recursos.

El primer principio defiende el uso de URIs para identificar no sólo documentos Web y contenidos digitales, sino también objetos del mundo real y conceptos abstractos, que deben estar representados en el formato RDF.

El segundo principio defiende el uso de URIs HTTP para identificar los objetos y los conceptos abstractos definidos por el Principio 1, lo que posibilita que dichos URI se puedan desreferenciar sobre un protocolo HTTP. En este contexto, desreferenciar es el proceso de recuperar una representación de un recurso identificado por un URI, donde un recurso puede tener varias representaciones como documentos HTML, RDF, XML u otros.

Para permitir que una amplia gama de aplicaciones pueda procesar los datos disponibles en la Web, es importante que exista un acuerdo sobre un formato estándar para disponibilizar los datos. El tercer principio de Datos Conectados defiende el uso de RDF como modelo para la publicación de datos estructurados en la Web (CYGANIAK; WOOD; LANTHALER, 2014). El RDF permite describir el significado de los recursos, habilitando agentes de software para explorar los datos de forma automática, a menudo agregando, interpretando o mezclando datos.

El cuarto principio se refiere al uso de enlaces para conectar no sólo los documentos de la Web, sino cualquier tipo de recurso. Por ejemplo, se puede crear un vínculo entre una persona y un lugar, o entre una ubicación y una empresa. En contraste con la Web clásica donde los hipervínculos son en gran parte no "tipeados", los hipervínculos que conectan los recursos en un contexto de Datos Conectados son capaces de describir la relación entre ellos. En el contexto de Datos Conectados, los hipervínculos se denominan vínculos RDF para diferenciarlos de los hipervínculos que existen en la Web convencional (HEATH; BIZER, 2011).

Es importante destacar que hoy en día ya existe un gran volumen de datos abiertos conectados disponible en la Web. Como ejemplo, se destacan los conjuntos de datos abiertos publicados por el proyecto LOD [8] .Como se mencionó anteriormente, los Datos Conectados contribuyen a generar una Web de Datos, por lo que constituyen la opción preferida para la publicación de datos en la Web. En este contexto, el W3C *Government Linked Data Working Group* propuso un conjunto de Buenas Prácticas para la publicación de Datos Conectados a fin de proveer directrices que faciliten el acceso y la reutilización de los datos gubernamentales abiertos [9].

http://lod-cloud.net [8]
http://www.w3.org/TR/ld-bp [9]

DATOS EN LA WEB

Datos en la Web es un término más general que se puede utilizar para referirse a los datos publicados de acuerdo con la base arquitectónica de la Web (JACOBS; WALSH, 2004). Como se ilustra en la Figura 2, los datos en la Web se pueden clasificar como Datos Abiertos (PIRES, 2015), Datos Conectados y Datos Abiertos Conectados (BERNERS-LEE, 2006). De acuerdo con el *Open Data Charter*, "datos abiertos son datos digitales disponibles con las características técnicas y jurídicas necesarias para que puedan ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, en cualquier momento y en cualquier lugar". Dado que la Web es el medio más adecuado para poner a disposición los datos abiertos, estos datos también son datos en la Web. Otra distinción importante se refiere a los datos publicados en la Web de acuerdo con los Principios de los Datos Conectados. Una parte de los datos actualmente disponibles en la Web sigue estos principios y se clasifica como Datos Conectados. Por último, cuando un conjunto de datos se publica en la Web siguiendo tanto los principios de los Datos Abiertos como los principios de los Datos Conectados, dichos datos se pueden clasificar como Datos Abiertos Conectados.

Es importante observar que no todos los conjuntos de datos publicados en la Web se comparten abiertamente, lo que significa que una gran parte de los datos publicados en la Web están "cerrados". A la hora de determinar la política de publicación de datos y en qué circunstancias deben publicarse los datos, quienes publican datos deben tener en cuenta la seguridad, la sensibilidad comercial y, sobre todo, la privacidad de las personas.

Figura 2:
Intersección de los Datos en la Web, Datos Abiertos y Datos Conectados
Fuente: Los autores

CICLO DE VIDA DE LOS DATOS EN LA WEB

El proceso de publicación y consumo de datos en la Web implica diferentes fases que van desde la selección y publicación de los datos hasta el uso de los datos y el *feedback* sobre los datos utilizados. Este conjunto de fases que componen el proceso de publicación y consumo de los datos se denomina Ciclo de Vida de los Datos en la Web. La Figura 3 muestra las fases del Ciclo de Vida de los Datos en la Web, las cuales se describen brevemente a continuación.

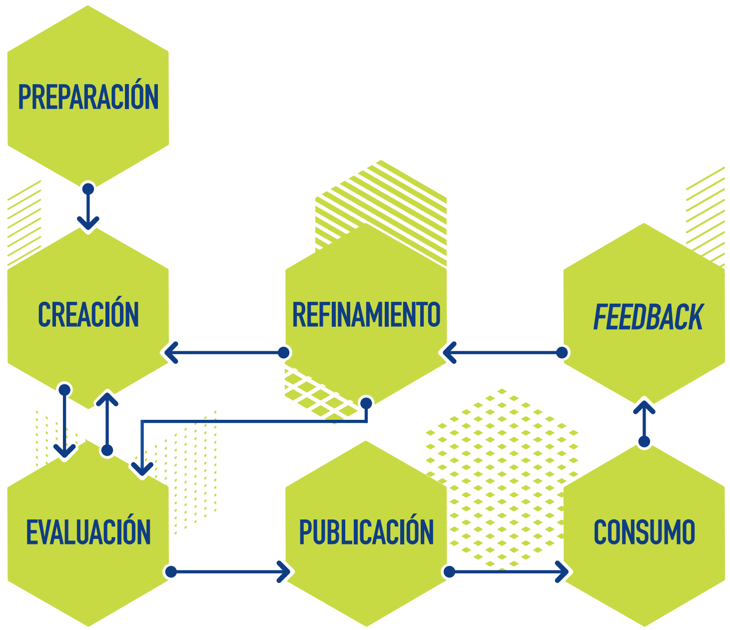


Figura 3:
Ciclo de vida de los datos en la Web
Fuente:
Los autores

- **Preparación:** Esta fase se extiende desde el momento en que surge la intención de publicar los datos hasta la selección de los datos que serán publicados. Cabe recordar que no existen reglas que determinen la prioridad de los datos a ser publicados, pero siempre es importante tener en cuenta la relevancia de los datos. En otras palabras, los datos que tienen un gran potencial de utilización deberían tener prioridad en el momento de la elección. De esta forma, siempre que sea posible, es importante consultar previamente a los potenciales consumidores de datos para identificar la relevancia de los datos.

■ **Creación:** Se refiere al momento en que se crean los datos, es decir, comprende la etapa que va desde la extracción de los datos de fuentes de datos ya existentes hasta su transformación al formato adecuado para su publicación en la Web. Durante la etapa de creación, además de los datos propiamente dichos, también se deben crear los metadatos que se usarán para describir a los datos. En la etapa de creación también se escogerán los formatos de datos que se utilizarán para publicar los datos y metadatos. Además, siempre es bueno considerar la publicación de los datos en diferentes formatos para así minimizar la necesidad de que los consumidores tengan que transformarlos.

■ **Evaluación:** Esta etapa se refiere a la evaluación de los datos antes de su publicación. Es importante que los expertos puedan evaluar los datos para detectar incoherencias o errores en los mismos, así como para señalar datos confidenciales que no deben publicarse, por ejemplo. Los datos solo deben estar disponibles para su publicación después de una cuidadosa evaluación. Cuando sea necesario, los datos pueden volver a la etapa anterior para resolver cualquier problema detectado por los expertos.

■ **Publicación:** Comprende el momento en que los datos se ponen a disposición del público en la Web. Para esto se pueden utilizar herramientas de catalogación de datos, como CKAN[10]y Sócrata.[11] También se pueden utilizar APIs (*Application Program Interfaces*) que permitan el acceso fácil a los datos publicados oa las páginas Web, por ejemplo. En todos los casos, el publicador de datos debe ofrecer toda la información necesaria para que el consumidor pueda acceder fácilmente a los datos. Además, es importante garantizar que los datos se actualizarán según una frecuencia predeterminada, la cual deberá estar disponible junto con los datos.

[10] http://ckan.org
[11] http://www.socrata.com

■ **Consumo:** Implica el momento en que los datos se utilizan para crear visualizaciones, como gráficos y mapas de calor, así como para aplicaciones que permiten cruzar y analizar los datos. Esta etapa del Ciclo de Vida está directamente relacionada con el consumidor de los datos, que puede ser desde una gran empresa interesada en utilizar los datos disponibles en la Web para mejorar sus productos y servicios, hasta un desarrollador interesado en emplear los datos para crear una aplicación que mejore la calidad de vida en su ciudad.

■ ***Feedback*:** Esta fase comprende el momento en que los consumidores deben aportar comentarios sobre los datos y metadatos previamente utilizados. Esta fase es de fundamental importancia, ya que el *feedback* de los consumidores permitirá identificar mejoras y realizar correcciones en los datos previamente publicados. Además, este canal de comunicación entre consumidores y publicadores de datos también facilita la identificación de nuevos datos relevantes que deben tener prioridad a la hora de escoger nuevos datos para su publicación.

■**Refinamiento:** Esta fase comprende todas las actividades relacionadas con las adiciones o actualizaciones de los datos que ya se han publicado. Es muy importante garantizar el mantenimiento de los datos previamente publicados, a fin de ofrecer mayor seguridad a quienes los consumirán. Se puede realizar un mantenimiento de acuerdo con el *feedback* de los consumidores, o bien se pueden generar nuevas versiones para garantizar que los datos no queden obsoletos. Para ello, es importante gestionar correctamente las diferentes versiones de los datos y garantizar que los consumidores tengan acceso a la versión correcta.

Con respecto a los actores que participan en el Ciclo de Vida de los Datos en la Web, estos pueden desempeñar dos papeles principales: los publicadores de datos y los consumidores de datos. El papel de publicador de datos puede ser desempeñado por varios actores, quienes son responsables de realizar actividades como la creación de metadatos, la creación y la publicación de datos. Los consumidores de datos son actores que reciben y consumen los datos. Es importante señalar que los consumidores de datos también pueden ser publicadores de datos, dado que los consumidores pueden mejorar y refinar los datos para ofrecerlos otra vez a la comunidad. También es importante observar que el Ciclo de Vida propuesto no requiere seguir todos los pasos antes de iniciar una nueva iteración.

BUENAS PRÁCTICAS PARA LOS DATOS EN LA WEB

Las Buenas Prácticas para los Datos en la Web (DWBP, del inglés *Data on the Web Best Practices*) descritas en la Recomendación del W3C de Lóscio, Burle y Calegari (2017) fueron desarrolladas para incentivar y permitir la expansión continuada de la Web como un medio para el intercambio de datos. En términos generales, quienes publican datos buscan compartirlos abiertamente o con acceso controlado. Los consumidores de datos quieren poder encontrar, utilizar y establecer conexiones entre los datos, especialmente si los datos son precisos, actualizados y tienen garantía de alta disponibilidad. Esto hace que sea fundamental la existencia de un entendimiento común entre los publicadores y los consumidores de datos. Sin este acuerdo, los esfuerzos de los publicadores podrían ser incompatibles con los deseos de los consumidores.

En este contexto, resulta fundamental ofrecer a los publicadores una orientación que pueda contribuir a mejorar la coherencia en la forma en que se gestionan los datos. Se espera que esta orientación promueva la reutilización de los datos y fomente la confianza en los datos por parte de los desarrolladores, cualquiera sea la tecnología que utilicen, para así aumentar el potencial de innovación genuina. El conjunto de Buenas Prácticas propuesto en Lóscio, Burle y Calegari (2017) fue desarrollo para ofrecer orientación técnica para la publicación de datos en la Web, contribuyendo a mejorar la relación entre publicadores y consumidores de datos.

Las Buenas Prácticas propuestas abarcan diferentes retos y requisitos relacionados con la publicación y el consumo de datos, entre ellos los formatos de datos, el acceso a datos identificadores de datos, los vocabularios y los metadatos. Por un lado, cada buena práctica se ocupa de al menos uno de los requisitos identificados en el documento de casos de uso en la Web (LEE; LÓSCIO; ARCHER, 2015), de tal forma que la relevancia de la buena práctica es evidente a partir de dichos requisitos. Por otro lado, cada requisito es abordado por al menos una buena práctica.

Como se describe en Lóscio, Burle y Calegari (2017) y se muestra en el Cuadro 1, para cada buena práctica hay un resultado esperado que describe "Lo que se debe poder hacer cuando un publicador de datos sigue la buena práctica". En general, el resultado esperado es una mejora en el modo en que un consumidor de datos (humano o software) puede manipular un conjunto de datos publicado en la Web. En algunos casos, el resultado esperado refleja una mejora en el propio conjunto de datos, lo que también beneficiará al consumidor de datos.

Las Buenas Prácticas propuestas para la publicación y utilización de datos en la Web se refieren a conjuntos de datos, es decir, "una colección de datos publicados, administrados por un único agente y disponibles para ser accedidos o descargados en uno o más formatos" (MAALI; ERICKSON, 2014, traducido por los autores). Con el término datos nos referimos a "hechos conocidos que pueden ser grabados y que tienen significado implícito" (ELMASRI; NAVATHE, 2010, traducido por los autores). Como se describe en la Figura 4, los datos se publican en diferentes distribuciones, que son una forma física específica de un conjunto de datos. Estas distribuciones facilitan el intercambio de datos a gran escala, lo que permite que los conjuntos de datos puedan ser utilizados por diferentes grupos de consumidores de datos. En otras palabras, "una persona o grupo accede, utiliza y potencialmente ejecuta las etapas de postratamiento de los datos" (STRONG; LEE; WANG, 1997, traducido por los autores), sin tener en cuenta su finalidad, destinatario, interés o licencia. Teniendo en cuenta esta heterogeneidad y el hecho de que los publicadores y consumidores de datos pueden no conocerse, es necesario proporcionar

cierta información sobre los conjuntos de datos y distribuciones que también pueda contribuir a su confiabilidad y reutilización, tales como: metadatos estructurales, metadatos descriptivos, acceso a la información, información sobre la calidad de los datos, información sobre su procedencia, información sobre la licencia e información sobre el uso.

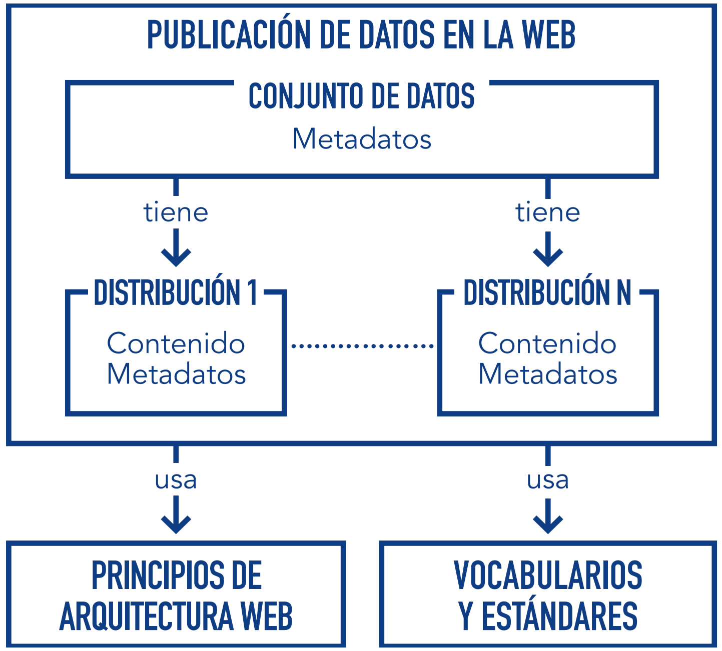


Figura 4: Contexto de publicación de datos en la Web. Fuente: Lóscio, Burle e Calegari (2017)

Por último, una cuestión importante sobre la publicación y el intercambio de datos en la Web tiene que ver con la base arquitectónica de la Web (JACOBS; WALSH, 2004). Aquí, un aspecto relevante es el principio de identificación, que indica que se deben utilizar URIs para identificar a los recursos. En nuestro contexto, un recurso puede ser un conjunto completo de datos o un elemento específico de un conjunto de datos determinado. Todos los recursos se deben publicar con URIs estables, de modo que puedan ser referenciados y que se puedan hacer conexiones entre dos o más recursos a través de URIs.

BUENAS PRÁCTICAS PARA LOS DATOS EN LA WEB CON SUS RESPECTIVOS RESULTADOS ESPERADOS

BP1
PROPORCIONAR METADATOS

Las personas podrán comprender los metadatos y los agentes de software podrán procesarlos.

BP2
PROPORCIONAR METADATOS DESCRIPTIVOS

Las personas podrán interpretar la naturaleza del conjunto de datos y sus distribuciones; los agentes de software podrán descubrir automáticamente conjuntos de datos y distribuciones.

BP3
PROPORCIONAR METADATOS ESTRUCTURALES

Las personas podrán interpretar el esquema de un conjunto de datos y los agentes de software podrán procesar automáticamente las distribuciones.

BP4
PROPORCIONAR INFORMACIÓN SOBRE LAS LICENCIAS DE LOS DATOS

Las personas podrán comprender la información sobre las licencias de los datos, describiendo eventuales restricciones impuestas al uso de ciertos datos; los agentes de software podrán detectar automáticamente la licencia de los datos de una distribución.

BP5
PROPORCIONAR INFORMACIÓN SOBRE LA PROCEDENCIA DE LOS DATOS

Las personas conocerán el origen de los conjuntos de datos y los agentes de software podrán procesar automáticamente la información de procedencia.

BP6
PROPORCIONAR INFORMACIÓN SOBRE LA CALIDAD DE LOS DATOS

Las personas y los agentes de software podrán evaluar la calidad y, por lo tanto, la adecuación de un conjunto de datos para su aplicación.

BP7
PROPORCIONAR UN INDICADOR DE VERSIÓN

Las personas y los agentes de software podrán determinar fácilmente la versión de un conjunto de datos.

BP8
PROPORCIONAR EL HISTORIAL DE VERSIONES

Las personas y los agentes de software podrán entender cómo el conjunto de datos cambia de una versión a otra y cómo difieren dos versiones específicas cualquiera.

BP9
UTILIZAR URIS PERSISTENTES COMO IDENTIFICADORES DE LOS CONJUNTOS DE DATOS

Los conjuntos de datos o la información sobre los conjuntos de datos podrán ser descubiertos y citados a lo largo del tiempo, independientemente de su disponibilidad o del formato de los datos.

BP10
UTILIZAR URIS PERSISTENTES COMO IDENTIFICADORES DENTRO DE LOS CONJUNTOS DE DATOS

Los elementos de datos se relacionarán en toda la Web, creando un espacio global de información accesible para las personas y las máquinas.

BP11
ASIGNAR URIS A LAS VERSIONES INDIVIDUALES DE LOS CONJUNTOS DE DATOS Y TAMBIÉN A LA SERIE COMPLETA

Las personas y los agentes de software podrán referirse a versiones específicas de un conjunto de datos, a una serie de conjuntos de datos, así como a la versión más reciente de un conjunto de datos.

BP12
UTILIZAR FORMATOS DE DATOS ESTANDARIZADOS LEGIBLES POR MÁQUINA

Las máquinas podrán leer y procesar los datos publicados en la Web y las personas podrán usar herramientas computacionales para manipular los datos.

BP13
UTILIZAR REPRESENTACIONES DE DATOS QUE SEAN INDEPENDIENTES DE LA CONFIGURACIÓN DE PARÁMETROS REGIONALES (*LOCALE NEUTRAL*)

Las personas y los agentes de software podrán interpretar el significado de las cadenas que representan fechas, horas, monedas y números con precisión.

BP14
PROPORCIONAR DATOS EN DIVERSOS FORMATOS

La mayor cantidad de usuarios posible podrá utilizar los datos sin antes tener que convertirlos a su formato preferido.

BP15
REUTILIZAR VOCABULARIOS, PREFERENTEMENTE VOCABULARIOS ESTANDARIZADOS

Se mejorará la interoperabilidad y el consenso entre los publicadores y los consumidores de datos.

BP16
ESCOGER EL NIVEL DE FORMALIZACIÓN ADECUADO

Los casos de aplicación más probables se admitirán sin más complejidad de la necesaria.

BP17
PROPORCIONAR DESCARGA MASIVA *'BULK DOWNLOAD'*

Se podrán realizar transferencias de grandes archivos, es decir, de archivos que requerirían más tiempo del que un usuario típico consideraría razonable, por medio de protocolos de transferencia de archivos dedicados.

BP18
PROPORCIONAR SUBCONJUNTOS PARA LOS GRANDES CONJUNTOS DE DATOS

Las personas y las aplicaciones podrán acceder a subconjuntos de un conjunto de datos antes que al conjunto completo, proporcionando a los consumidores el acceso a los datos con una gran proporción de datos que son realmente necesarios en comparación con los datos innecesarios. Los conjuntos de datos estáticos considerados demasiado grandes se podrán descargar en partes más pequeñas. Se podrán utilizar APIs para filtrar los datos disponibles. La granularidad de los datos se podrá definir de acuerdo con las necesidades del dominio y las demandas de desempeño de las aplicaciones.

BP19
USAR NEGOCIACIÓN DE CONTENIDO PARA SERVIR LOS DATOS DISPONIBLES EN DIFERENTES FORMATOS

El uso de negociación de contenidopermitirá que se puedan servir diferentes recursos o diferentes representaciones de un mismo recurso según la solicitud enviada por el cliente.

BP20
PROPORCIONAR ACCESO EN TIEMPO REAL

Las aplicaciones podrán acceder a los datos en tiempo real o casi en tiempo real, considerando que "en tiempo real significa en un intervalo que va de milisegundos hasta algunos segundos después de la creación de los datos."

BP21
PROPORCIONAR DATOS ACTUALIZADOS

Los datos en la Web se actualizarán de manera oportuna para que los datos disponibles en línea reflejen los datos más recientes divulgados a través de cualquier otro canal. Cuando haya nuevos datos disponibles, estos se publicarán en la Web tan pronto como sea posible.

BP22
PROPORCIONAR UNA EXPLICACIÓN PARA LOS DATOS QUE NO ESTÁN DISPONIBLES

Los consumidores sabrán que los datos a los que se hace referencia a partir del conjunto de datos no están disponibles o que están disponibles bajo otras condiciones.

BP23
HACER QUE LOS DATOS ESTÉN DISPONIBLES A TRAVÉS DE UNA API

Los desarrolladores tendrán acceso a los datos para utilizarlos en sus propias aplicaciones, con datos actualizados y sin que se requiera ningún esfuerzo por parte de los consumidores. Las aplicaciones podrán obtener datos específicos mediante consultas a la API.

BP24
UTILIZAR ESTÁNDARES WEB COMO BASE PARA LA CONSTRUCCIÓN DE APIS

Los desarrolladores que tengan alguna experiencia con las APIs basadas en estándares Web (por ejemplo, REST) ya tendrán un conocimiento inicial de cómo utilizar la API. Además, el mantenimiento de las API será más sencillo.

BP25
PROPORCIONAR DOCUMENTACIÓN COMPLETA PARA LAS APIS

Los desarrolladores podrán obtener información detallada sobre cada llamada a la API, incluidos los parámetros que lleva y lo que se espera que debe retornar, es decir, todo el conjunto de información relacionada con la API. El conjunto de valores –cómo usarlo, notificaciones de cambios recientes, información de contacto, etc. – debe estar descrito y ser fácilmente navegable en la Web. Esto también permite que las máquinas accedan a la documentación de la API para ayudar a los desarrolladores a crear software de cliente API.

BP26
EVITAR CAMBIOS QUE AFECTEN EL FUNCIONAMIENTO DE SU API

El código del desarrollador deberá seguir funcionando después de cambios en la API. Los desarrolladores deberán estar al tanto de las mejoras que se realizan en la API y deberán poder hacer uso de ellas. Los cambios que afecten el funcionamiento de la API serán poco frecuentes y, si ocurren, los desarrolladores tendrán tiempo e información suficiente para adaptar su código. Esto les permitirá evitar fallas y aumentar la confianza en la API. Los cambios en la API se deberán anunciar en el sitio Web de la documentación de dicha API.

BP27
PRESERVAR IDENTIFICADORES

El URI de un recurso siempre hará referencia al conjunto de datos o redireccionará a información sobre el mismo.

BP28
EVALUAR LA COBERTURA DEL CONJUNTO DE DATOS

Los usuarios podrán hacer uso de los datos archivados en el futuro.

BP29
RECOLECTAR *FEEDBACK* DE LOS CONSUMIDORES DE DATOS

Los consumidores de datos podrán ofrecer *feedback* y calificaciones sobre los conjuntos de datos y distribuciones.

BP30
COMPARTIR EL *FEEDBACK* DISPONIBLE

Los consumidores podrán evaluar los tipos de errores que afectan al conjunto de datos, revisar las experiencias de otros usuarios y estar seguros de que quien publica los datos trata los problemas de forma adecuada. Los consumidores también podrán determinar si otros usuarios ya han realizado comentarios similares, ahorrándoles la molestia de presentar informes de *bugs* innecesarios y evitando que los publicadores tengan que lidiar con duplicados.

BP31
ENRIQUECER LOS DATOS MEDIANTE LA GENERACIÓN DE NUEVOS DATOS

Los conjuntos de datos con valores nulos se podrán "corregir" mediante el llenado de dichos valores. Se conferirá estructura a los datos y su utilidad se mejorará si se agregan medidas o atributos relevantes, pero solo si la adición no altera los resultados analíticos, el significado o el poder estadístico de los datos.

BP32
PROPORCIONAR VISUALIZACIONES COMPLEMENTARIAS

Complementar los conjuntos de datos con posibles visualizaciones permitirá que los consumidores humanos tengan una visión inmediata de los datos, presentándolos de formas fácilmente comprensibles.

BP33
PROVEER *FEEDBACK* AL PUBLICADOR ORIGINAL

Una mejor comunicación entre publicadores y consumidores hará que sea más fácil para los publicadores originales determinar cómo se están utilizando los datos que publican, lo que a su vez les ayudará a justificar la publicación de los datos. A los publicadores también se les informará de las medidas que pueden tomar para mejorar sus datos. Esto contribuye a una mejora de los datos en general.

BP34
OBEDECER LOS TÉRMINOS DE LAS LICENCIAS

Quienes publican datos podrán confiar que su trabajo se está reutilizando de acuerdo con sus requisitos de licenciamiento, lo que aumentará la probabilidad de que continúen publicando datos. Quienes reutilizan dichos datos también podrán licenciar adecuadamente sus trabajos derivados.

BP35
CITAR LA PUBLICACIÓN ORIGINAL DEL CONJUNTO DE DATOS

Los usuarios finales podrán evaluar la confiabilidad de los datos que ven y se reconocerán los esfuerzos de los publicadores originales. La cadena de procedencia de los datos en la Web se podrá rastrear hasta su publicador original.

Para animar a los publicadores a adoptar estas Buenas Prácticas para la publicación de datos en la Web, se identificó una serie de beneficios que se pueden lograr a partir de la aplicación de las mismas: comprensión, facilidad de procesamiento, facilidad de descubrimiento, reutilización, confianza, capacidad de conexión de datos, facilidad de acceso e interoperabilidad. Estos beneficios son importantes porque ayudan a los publicadores de datos a entender mejor "lo que será posible" una vez que se adopten las buenas prácticas. Cada beneficio está asociado a una o más Buenas Prácticas. Por ejemplo, la "comprensibilidad" está asociada con diez Buenas Prácticas, que están relacionadas con los metadatos, los vocabularios de datos, el *feedback* y el enriquecimiento de los datos. Esto significa que, si un publicador de datos adopta estas prácticas, aumentará el nivel de comprensibilidad, es decir, las personas comprenderán mejor la estructura y el significado de los datos, así como la naturaleza del conjunto de datos. Cabe destacar que el beneficio se refuerza a medida que aumenta la adopción de las Buenas Prácticas. Considerando que la publicación de datos en la Web es un proceso incremental, el nivel de cada beneficio podrá aumentar después de algunas iteraciones del proceso de publicación de datos.

- **Comprensión:** Las personas tendrán una mejor comprensión de la estructura y del significado de los datos, así como de los metadatos y de la naturaleza del conjunto de datos.

- **Facilidad de procesamiento:** Las máquinas o los agentes de software podrán procesar y manipular automáticamente los datos.

- **Facilidad de descubrimiento:** Los agentes de software podrán descubrir automáticamente un conjunto de datos o datos dentro de un conjunto de datos.

- **Reutilización:** Las chances de que diferentes grupos de consumidores de datos reutilicen el conjunto de datos tenderán a aumentar.

- **Confianza:** La confianza de los consumidores en el conjunto de datos tenderá a mejorar.

- **Capacidad de conexión:** Será posible crear enlaces entre los conjuntos de datos y los elementos de datos.

- **Facilidad de acceso:** Las personas y las máquinas podrán acceder a datos actualizados de diferentes maneras.

- **Interoperabilidad:** Será más fácil llegar a un consenso entre los publicadores y los consumidores de datos.

TÉCNICAS PARA LA PUBLICACIÓN DE DATOS EN LA WEB

A medida que la Web se fue consolidando como una plataforma para la publicación y el intercambio de documentos, las organizaciones comenzaron a tener interés en utilizar la Web como plataforma para la publicación de datos. En los últimos años han surgido diferentes técnicas para la publicación de datos en la Web que van desde el uso de formularios para realizar consultas a una base de datos hasta la publicación de Datos Conectados (CERI et al., 2013 y FERRARA et al., 2014). A continuación, se presentan algunas de estas técnicas para la publicación de datos (CERI et al., 2013 y FERRARA et al., 2014), entre ellas el uso de APIs Web, la inserción de datos directamente en las páginas HTML y las herramientas para la creación de catálogos de datos.

ACCESO A PARTIR DE APIS WEB

Una forma de publicar datos en la Web consiste en utilizar APIs Web. Una de las primeras propuestas para la estandarización de las APIs para la Web fueron los *Web Services* (ALONSO et al., 2004), inspirados en el paradigma de RPC (*Remote Procedure Call*) (NELSON, 1981) y el uso de XML (*eXtensible Markup Language*) para el intercambio de datos. Posteriormente, surgió el paradigma REST (*REpresentational State Transfer*) y el formato JSON (*JavaScript Object Notation*) (MANDEL, 2008) pasó a ser ampliamente adoptado. Este nuevo tipo de API se conoce como servicio RESTful.

En general, los datos expuestos por medio de APIs no pueden ser encontrados por los motores de búsqueda. Una de las razones de lo anterior es que en muchos casos es necesario realizar una autenticación antes de poder acceder a la API. Además, existen restricciones en cuanto al uso de la API para evitar accesos exhaustivos a los datos. Por lo tanto, se puede decir que los datos disponibles por medio de APIs son similares a los datos disponibles en la *Deep Web*, es decir, no se pueden encontrar e indexar fácilmente. Sin embargo, esto se debe a una razón muy diferente que consiste en la necesidad de los publicadores de controlar el acceso a los datos por parte de las aplicaciones externas.

ENRIQUECIMIENTO DE PÁGINAS HTML

Otra forma de publicar datos en la Web consiste en incluir los datos en las páginas HTML. Esto se puede hacer usando microformatos, es decir, marcadores (*tags*) específicos que explicitan la semántica de los datos. El uso de microformatos permite a los motores de búsqueda identificar los datos disponibles en las páginas HTML y así presentar mejores resultados a los usuarios. Además, los publicadores de datos pueden alcanzar una mayor visibilidad. La comunidad ha desarrollado diversos microformatos para la publicación de datos de diferentes dominios, entre ellos: *hCalendar* para eventos, *hReview* para revisiones y *ratings*, *hRecipe* para recetas culinarias y *hCard* para datos personales[12].

El uso de microformatos es una solución simple para la publicación de datos en la Web, aunque también tiene algunas limitaciones: I) el uso de diferentes microformatos en una misma página puede llevar a conflictos de nombres (por ejemplo, la *class url* de CSS y el término url del microformato *hCalendar*), II) no permite crear especializaciones y generalizaciones, y III) cada microformato requiere un *parser* específico.

Estos problemas se pueden solucionar utilizando RDFa,[13] una solución que permite especificar atributos para la descripción de datos estructurados en cualquier lenguaje de marcado, en particular XHTML[14] y HTML. Mientras que los microformatos combinan la sintaxis para incluir los datos estructurados en las páginas HTML con la propia semántica de los datos, RDFa solo se preocupa de la sintaxis para la inclusión de los datos estructurados. Para la semántica de los datos, RDFa permite el uso de vocabularios específicos, como el schema.org[15]. RDFa permite utilizar juntos múltiples vocabularios sin la necesidad de *parsers* específicos para cada uno de ellos.

Además de utilizar RDFa para agregar metadatos estructurados en un documento HTML, también se puede utilizar el lenguaje JSON-LD[16] (JSON *for Linked Data*). Se trata de un estándar basado en el formato JSON, pero que también permite el uso de vocabularios y ontologías para la descripción de los datos. El formato JSON-LD tiene una gran adopción dentro de la comunidad técnica y Google[17] recomienda su adopción como formato estándar para el intercambio de Datos Conectados en las páginas Web.

HERRAMIENTAS PARA CATALOGACIÓN DE DATOS

Con el creciente interés en la publicación de datos abiertos, particularmente datos abiertos gubernamentales, se ha destacado una nueva forma de publicar datos en la Web: las herramientas para catálogos de datos, por ejemplo, CKAN[18] y Socrata[19]. A partir de estas plataformas se crean los portales de datos abiertos, que ofrecen acceso a conjuntos de datos previamente catalogados. Los conjuntos de datos se organizan como una serie de recursos y se pueden clasificar de acuerdo con etiquetas (*tags*) que explicitan el dominio de los datos.

Los portales de datos son una excelente herramienta para indexar conjuntos de datos, pero dejan que desear en cuanto a la búsqueda de datos, ya que no permiten realizar búsquedas en los conjuntos de datos propiamente dichos. En algunos casos, las herramientas de catalogación ofrecen APIs de acceso a los datos, aunque esto se hace de forma bastante simplificada. Los conjuntos de datos disponibles en los catálogos pueden ser encontrados por las herramientas de búsqueda, pero aún no es posible encontrar elementos de datos específicos almacenados en un conjunto de datos.

A pesar de la gran difusión de los portales de datos abiertos, estas soluciones tienen diferentes limitaciones, entre las cuales se destacan: la dificultad para mantener los datos actualizados, la falta de estándares de metadatos para la descripción de los conjuntos de datos y la imposibilidad de realizar consultas sobre los datos. Además, como los conjuntos de datos publicados en los portales generalmente están disponibles en diferentes formatos, es decir, hay múltiples archivos para un mismo conjunto de datos, también puede haber redundancia.

12 http://microformats.org
13 http://w3.org/TR/rdfa-primer
14 http://w3.org/TR/xhtml1
15 http://schema.org
16 https://www.w3.org/TR/json-ld
17 https://developers.google.com/search/docs/guides/intro-structured-data
18 http://ckan.org
19 http://www.socrata.com

CONCLUSIÓN

El interés en la publicación de datos en la Web no es nuevo. Sin embargo, el creciente interés en el uso de la Web como una plataforma para compartir datos trae nuevos desafíos para la publicación de datos en forma estructurada. En escenarios donde los consumidores de datos no se conocen en forma anticipada, la publicación de datos se debe realizar de manera de satisfacer a grupos de consumidores con diferentes requisitos y perfiles.

En este contexto, además de los aspectos básicos de la disponibilidad de los datos, también es necesario tener en cuenta otros aspectos que se refieren a la comprensión, la confiabilidad y el procesamiento automático de los datos. Por un lado, los publicadores deben proporcionar información que ayude a entender los datos, como por ejemplo metadatos estructurales, pero también deben proporcionar información que permita a los consumidores conocer la procedencia y la calidad de los datos. Por otro lado, los consumidores deben poder proveer *feedback* sobre los datos utilizados, para así contribuir a mejorar el proceso de publicación. Además, los consumidores deben proporcionar información sobre el uso de los datos, es decir, junto con la aplicación o la visualización generada a partir de los datos publicados, se debe facilitar información sobre los datos que se utilizaron. Para facilitar las tareas de los publicadores y consumidores de datos en la Web, se han propuesto un conjunto de Buenas Prácticas que abordan aspectos relacionados a todo el Ciclo de Vida de los Datos en la Web. La adopción de estas Buenas Prácticas lleva a la creación un canal de comunicación entre proveedores y consumidores, además de contribuir a la mejora del proceso de publicación de datos en la Web.

REFERENCIAS

ABITEBOUL, Serge; BUNEMAN, Peter; SUCIU, Dan. Data on the Web: from relations to semistructured data and XML. San Francisco: Morgan Kaufmann, 2000.

ALONSO, Gustavo et al. Web Services: Concepts, Architectures and Applications. Heidelberg: Springer, 2004.

BERNERS-LEE, Tim; CONNOLLY, Dan; SWICK, Ralph R. Web Architecture: Describing and Exchanging Data. 1999. Disponible en: <https://www.w3.org/1999/04/WebData>. Accedido el 04 de septiembre de 2018.

BERNERS-LEE, Tim. Linked Data. 2006. Disponible en: <https://www.w3.org/DesignIssues/LinkedData.html>. Accedido el 04 de septiembre de 2018.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, v. 5, n. 3, p.1-22, jul. 2009. IGI Global.

CERI, Stefano et al. Web Information Retrieval. Springer Science & Business Media, 2013.

CYGANIAK, Richard; WOOD, David; LANTHALER, Markus. RDF 1.1 Concepts and Abstract Syntax. 2014. Disponible en: <https://www.w3.org/TR/rdf11-concepts/>. Accedido el 04 de septiembre de 2018.

ELMASRI, Ramez; NAVATHE, Shamkant. Fundamentals of Database Systems. Addison-wesley Publishing Company, 2010.

FERRARA, Emilio et al. Web data extraction, applications and techniques: A survey. Knowledge-based Systems, [s.l.], v. 70, p.301-323, nov. 2014. Elsevier BV.

GOLDSTEIN, Brett; DYSON, Lauren (Ed.). Beyond Transparency: Open Data and the Future of Civic Innovation. San Francisco: Code for America Press, 2013.

HEATH, Tom; BIZER, Christian. Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool Publishers, 2011. 136 p. (Synthesis Lectures on the Semantic Web: Theory and Technology).

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. Dados abertos conectados. San Pablo Paulo: Novatec, 2015. 175 p.

JACOBS, Ian; WALSH, Norman. Architecture of the World Wide Web. 2004. Disponível em: <https://www.w3.org/TR/webarch/>. Accedido el 04 de septiembre de 2018.

LEE, Deirdre; LÓSCIO, Bernadette Farias; ARCHER, Phil. Data on the Web Best Practices Use Cases & Requirements. 2015. Disponible en: <https://www.w3.org/TR/dwbp-ucr/>. Accedido el 04 de septiembre de 018.

LÓSCIO, Bernadette Farias; BURLE, Caroline; CALEGARI, Newton. Data on the Web Best Practices. 2017. Disponible en: <https://www.w3.org/TR/dwbp/>. Accedido el 04 de septiembre de 2018.

MAALI, Fadi; ERICKSON, John. Data catalog vocabulary (DCAT). 2014. Disponível em: <https://www.w3.org/TR/vocab-dcat/>. Accedido el 04 de septiembre de 2018.

NELSON, Bruce Jay. Remote procedure call. 1981. 201 f. Tesis (Doctorado) - School of Computer Science, Carnegie Mellon University, Pa, 1981.

OPEN KNOWLEDGE. Open data handbook. 2012. Disponível em: <http://opendatahandbook.org/>. Acesso em: 04 set. 2018.

PIRES, Marco Túlio. Guia de Dados Abertos. São Paulo: Este Guia é parte integrante do Projeto de Cooperação entre o Governo do Estado de São Paulo e o Reino Unido, 2015. Disponible en: <http://ceweb.br/media/docs/publicacoes/13/Guia_Dados_Abertos.pdf>. Accedido el 04 de septiembre de 2018.

STRONG, Diane M.; LEE, Yang W.; WANG, Richard Y. Data quality in context. Magazine Communications of the ACM, Nueva York, v. 40, n. 5, p.103-110, 05 de mayo de 1997.

TAUBERER, Joshua; LESSIG, Larry. The 8 Principles of Open Government Data. 2007. Disponible en: <https://opengovdata.org/>. Accedido el 04 de septiembre de 2018

ANEXO

HOJA DE RUTA PARA LA PUBLICACIÓN DE DATOS ABIERTOS

1.PREPARACIÓN

¿QUÉ HACER?	¿CÓMO HACERLO?	ARTEFACTOS	METADATOS
Identificar demandas de datos	**1.** Interactuar con potenciales consumidores a través de entrevistas o consultas públicas **2.** Analizar las solicitudes de acceso a la información **3.** Evaluar portales corporativos u otras fuentes de diseminación de datos	Plan de demandas de datos	
Identificar potenciales conjuntos de datos	**1.** Agrupar las demandas que se refieren a elementos de datos similares en un mismo conjunto de datos	Lista de conjuntos de datos	Descriptivos
Definir la prioridad de los conjuntos de datos a ser abiertos	**1.** Definir la prioridad de la apertura de cada conjunto de acuerdo con el número de solicitantes de cada demanda	Lista de prioridades para la apertura	

2.CREACIÓN

¿QUÉ HACER?	¿CÓMO HACERLO?	ARTEFACTOS	METADATOS
Modelado del conjunto de datos	**1.** Evaluar las propiedades de cada demanda asociada al conjunto de datos para definir la estructura del conjunto como un todo **2.** Agrupar las propiedades similares, eliminar las propiedades redundantes	Esquema inicial del conjunto de datos	Estructurales
Identificar las fuentes de datos de origen	**1.** Evaluar los sistemas y documentos existentes para identificar la fuente de origen de los datos	Lista de fuentes de datos de origen	Procedencia
Mapeo entre las fuentes de origen y el conjunto de datos	**1.** Establecer la correspondencia entre las propiedades del esquema del conjunto de datos y las propiedades de las fuentes de datos de origen	Documento de mapeo entre la fuente y el conjunto de datos	Descriptivos
Identificar los datos sensibles	**1.** Consultar con especialistas o recurrir a la legislación correspondiente para identificar los datos sensibles	Lista de datos sensibles	
Identificar vocabularios	**1.** Evaluar el uso de vocabularios conocidos en la definición de las propiedades del conjunto de datos (p. ej.: dcterms, foaf, schema.org) **2.** Buscar en repositorios de vocabularios para identificar vocabularios adecuados para el dominio	Lista de vocabularios que se utilizarán en el esquema del conjunto	

2.CREACIÓN

¿QUÉ HACER?	¿CÓMO HACERLO?	ARTEFACTOS	METADATOS
Mapeo entre los vocabularios y el esquema del conjunto de datos	**1.** Establecer la correspondencia entre las propiedades del esquema del conjunto de datos y los términos de los vocabularios previamente escogidos	Documento de mapeo entre el esquema y los vocabularios	
Definir la estrategia para la extracción de datos	**1.** De acuerdo con el tipo de fuente de datos (p. ej., base de datos, hoja de cálculo, documento de texto), especificar la forma de extracción de los datos	Plan de extracción de datos	Procedencia
Definir subconjuntos de datos	**1.** Si el volumen de datos es muy grande, definir posibles conjuntos de datos **2.** La división de los subconjuntos se puede realizar, por ejemplo, sobre la base de algún atributo temporal o espacial. También se pueden utilizar otros atributos más específicos	Lista de subconjuntos de conjuntos de datos	
Generar distribuciones	**1.** Aplicar una estrategia de extracción previamente definida y generar las distribuciones de datos deseadas	Distribuciones del conjunto de datos	Descriptivos de las distribuciones

4.PUBLICACIÓN

¿QUÉ HACER?	¿CÓMO HACERLO?	ARTEFACTOS	METADATOS
Publicar el conjunto de datos en una herramienta de catalogación de datos	**1.** El procedimiento puede variar según la herramienta utilizada. En general, es necesario cargar los archivos de las distribuciones y los metadatos del conjunto de datos **2.** Completar todos los metadatos solicitados y añadir nuevos metadatos si fuera necesario	Conjunto de datos disponible para acceso y descarga en la herramienta de catalogación	Descriptivos, Versionado
Publicar el conjunto de datos en una página HMTL	**1.** Crear la página HTML, tanto en una versión para consumo humano como en una versión para procesamiento por máquinas **2.** Insertar etiquetas RDFa en el código HTML con la información semántica para el procesamiento por máquinas	Conjunto de datos disponible para acceso y descarga en una página HTML	Descriptivos, Versionado
Desarrollar una API de acceso a los datos	**1.** Crear una API que permita acceder a los conjuntos de datos **2.** Crear la documentación de la API	Conjunto de datos disponible para acceso y descarga a través de una API y documentación de la API	Descriptivos, Versionado
Establecer un canal de comunicación con los consumidores de datos	**1.** El canal de comunicación dependerá de cómo se publicó el conjunto de datos. Si la herramienta utilizada no ofrece un canal de comunicación, crear una página HMTL	Página de contacto	Uso dos dados

3.EVALUACIÓN

¿QUÉ HACER?	¿CÓMO HACERLO?	ARTEFACTOS	METADATOS
Evaluar la calidad de los datos	**1.** Definir los criterios de calidad a ser evaluados (p. ej., «completitud», corrección, actualidad) **2.** Definir métricas para la evaluación de los criterios **3.** Definir requisitos mínimos para cada criterio de calidad **4.** Evaluar los criterios de calidad en forma manual o automática	Documento de calidad de los datos	Calidad de los datos
Liberar los datos para su publicación	**1.** Completar el documento de liberación del conjunto de datos	Documento de liberación del conjunto de datos	Descriptivos
Devolver el conjunto de datos a la fase de creación	**1.** Completar un documento de devolución a la fase de creación con la debida justificación y una descripción de las mejoras necesarias	Documento de devolución a la fase de creación