Linked Data: qué sucede con la heterogeneidad y la interoperabilidad

Linked Data: what happens to heterogeneity and interoperability

Jhon Francined Herrera-Cubides[1], Paulo Gaona-García[1], Salvador Sánchez-Alonso[2]
[1]*Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.*
[2]*Departamento de Ciencias de la Computación, Universidad de Alcalá de Henares, España.*
jfherrerac@udistrital.edu.co, pagaonag@udistrital.edu.co,
salvador.sanchez@uah.es.

***Resumen*— Linked Data – LD, permite construir la Web de los datos, fundamentada en la aplicación de principios básicos que contribuyen al crecimiento de la Web. En este proceso, LD enfrenta diferentes escenarios que parecen limitar el proceso de vinculación realizado. Este artículo expone las problemáticas de heterogeneidad, interoperabilidad y calidad de los datos, como factores que restringen el proceso de vinculación de los datos. Para ello, el diseño metodológico realiza una exploración bibliográfica, los resultados examinan y discuten las posturas acerca de las problemáticas citadas, y la posible incidencia de estas en la calidad de los datos enlazados. Como conclusión se plantea que hablar de heterogeneidad es intrínseco, puesto que persé la información siempre será heterogénea. Los modelos de datos pueden ser distintos pero si el formato de datos es común, la interoperabilidad es posible. Obviamente los modelos de datos deberían ser conocidos y públicos, ya que de no serlo existirá un problema clave que impide el desarrollo de la Web de los Datos**

***Palabras clave*— Linked Data, RDF, Interoperabilidad, Vocabulario, Heterogeneidad, Calidad de Datos.**

***Abstract*— Linked Data – LD, allows the construction of the Web of Data, which use basic principles that contribute to the growth of the Web. The process developed by LD faces different scenarios that seem to limit the linking process. This paper exposes the problems of heterogeneity, interoperability and data quality, as factors that restrict the process of data linkage. For this, the methodological design performs a bibliographic exploration; the results examine and discuss the postures about the mentioned problems, and the possible incidence of these in the quality of the linked data. In conclusion, speaking about heterogeneity is intrinsic, since the information will always be heterogeneous. Data models may be different but if the data format is common, interoperability is possible. Data models must be known and public, in order to avoid problems that prevent the development of the Web Data.**

***Key Word* — Linked Data, RDF, Interoperability, Vocabulary, Heterogeneity, Data Quality.**

I. INTRODUCCIÓN

Linked Data – LD, como uno de los conceptos claves de la Web Semántica, ofrece un conjunto de principios para compartir (exponer y consumir) datos vinculados en la Web. Esta estrategia hace uso de tecnologías como HTTP [1], RDF [2], XML [3], Sparql, URIs [4], entre otras, que permiten procesar y compartir información de manera comprensible por las máquinas. En 2009, Berners-Lee presentó los Principios de LD [5], los cuales enmarcan un esquema de 5 niveles que los datos deben cumplir para poderse vincular. El cambio de nivel en dicha jerarquía, requiere de cumplir una serie de requerimientos, con el fin de lograr una exposición adecuada, poder ser localizados y posteriormente ser consumidos. Bajo esta jerarquía, la vinculación de datos ha tenido un crecimiento gradual y significativo como se observa en los diferentes grafos de conocimiento [6] expuesto en la Web (Figura 1).

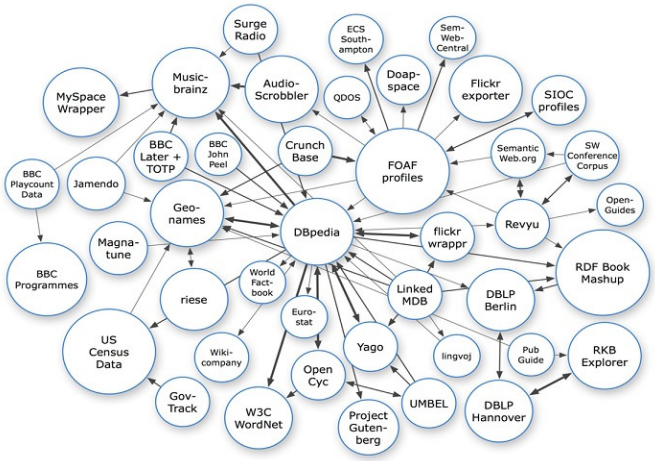

Fig. 1. Ejemplo Grafo de Conocimiento. Fuente: [7]

Adicional a ello, la tecnología de la Web Semántica (LD, OWL, RDF) combinada con técnicas de "Machine Learning" permiten crear agentes inteligentes que pueden ampliar su base

Fecha de Recepción: 20 de octubre de 2018
Fecha de Aceptación: 25 de junio de 2018

de conocimiento a todo tipo de contenidos semánticos enlazados en Internet, así como aprender a través de la interacción con su ”entrenador” y otras personas que interaccionen con él [8]. No obstante, LD se ve enfrentada a desafíos y retos que, de alguna forma, han influenciado en su aplicación y desarrollo en diferentes dominios de conocimiento. Dentro de estos problemas se encuentra la interoperabilidad y heterogeneidad de fuentes, de vocabularios y de tecnologías, y que circundan la calidad de los datos, aspectos que inciden en el grado de confianza ofrecido a los consumidores de recursos. Para contextualizar en este tema, [9] plantea el siguiente problema:

“Dentro de un sitio OCW existe un gran número de recursos educativos a los cuales no se puede acceder fácilmente ya que la mayoría de estos sistemas emplean sus propios modelos de datos para la descripción de sus recursos educativos, por esta razón las tareas de búsqueda, recuperación e integración se vuelven temas difíciles de alcanzar. Nace así la idea de la interoperabilidad entre sitios heterogéneos”.

Pero, ¿cómo se podría entender la heterogeneidad y la interoperabilidad, y su incidencia en la calidad de los datos?. Para comprender un poco más estos factores, el presente artículo se orienta a explorar, de forma teórica, conceptos claves identificados alrededor de dichos aspectos. En primera instancia, se contextualizara, de forma breve, una base del lenguaje usado en LD, para ubicar al lector en el área de conocimiento; posteriormente se conceptualizarán los referentes identificados que se involucran en la temática de heterogeneidad, interoperabilidad y calidad de los datos.

II. METODOLOGÍA

La presente propuesta se basa en una metodología de investigación de tipo descriptiva, donde se busca analizar qué es y cómo se manifiestan las variables de heterogeneidad e interoperabilidad en la vinculación de datos, y su relación con la calidad misma de los datos vinculados. A partir de este panorama, se identifica la necesidad de llevar a cabo una exploración documental con el fin de proveer información y poder establecer criterios de reflexión acerca de ¿Cómo la heterogeneidad y la interoperabilidad influyen en la calidad de los datos vinculados? Para llevar a cabo esta propuesta, a continuación, se definen un diseño metodológico estructurado en fases, que permiten determinar los procesos conducentes a nuestra propuesta de investigación.

Para llevar a cabo la metodología descriptiva, el diseño metodológico plantea las siguientes etapas:

a. Formulación de la pregunta de investigación: Para esta investigación se buscó plantear una discusión acerca de ¿Cómo la heterogeneidad y la interoperabilidad influyen en la calidad de los datos vinculados?
b. Revisión documental: se consultaron diferentes fuentes bibliográficas, tales como IEEE, Scopus, ACM, entre otros, definiendo criterios de búsqueda: LD, Linked Open Data, Heterogeneidad, Interoperabilidad, Calidad de Datos.
c. Análisis de recursos: Seguidamente se procedió a clasificarlos, y a analizar la información obtenido.
d. Síntesis de la información: Se abstrajo la información pertinente para documentar las variables planteadas.
e. Planteamiento de discusión y reflexión: se planteó la discusión acerca de las temáticas propuestas, y se construyeron las respectivas conclusiones.

III. DESARROLLO DEL DISEÑO METODOLÓGICO

A. Principios Básicos de Linked Data

LD se fundamenta en cuatro principios [10] básicos de vinculación de datos, los cuales se ilustran en la Figura 2 y son descritos en la tabla 1.

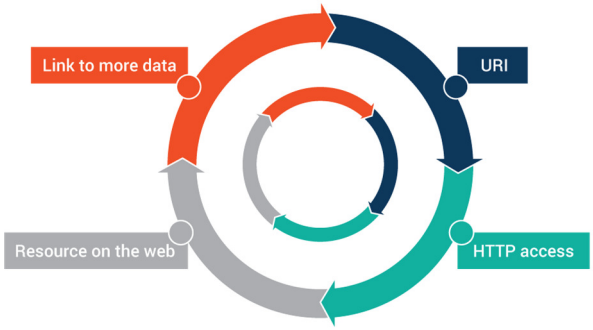

Fig. 2. Principios de LD. Fuente: [11]

Principio 1: Use URI (Identificador Universal de Recursos), como nombres de las cosas.	Identificación de las cosas. Si no se puede identificar una cosa, no se puede hablar de ello. Las URI se usan para nombrar cosas en LD son una versión generalizada de las URL que se utilizan para localizar páginas web a través del navegador
Principio 2: Usar HTTP URI para que la gente puede buscar esos nombres	Al escribir esa URI en el navegador, este le informará que no sabe cómo manejar la situación, ya que no es un tipo de URI que implementa. De ahí que es necesario que las URI sean resolubles en la Web, por ende se usa HTTP URI.
Principio 3: Cuando se usa una URI, esta debe proporcionar información útil.	Cualquier HTTP URI se puede escribir en un navegador Web y el navegador sabrá qué hacer con ella (por ejemplo: determinar el número de host, el puerto a utilizar, etc.). Si el servidor remoto responde afirmativamente, devolverá una representación del recurso en

	diferentes formatos como RDF, entre otros. De cualquier forma, se desearía que las URI puedan resolver unas descripciones útiles acerca de lo que usted ha nombrado.
Principio 4: Incluir enlaces a otros URI.	Los datos son más útiles si se vinculan datos relacionados, documentos y descripciones. Como se utilizó HTTP URI para publicar sus datos, otras personas pueden vincular sus datos. La capacidad de seguir estos enlaces le permite a la gente navegar por la Web de Datos igual que pueden navegar por la Web de Documentos.

Tabla I. Principios de LD. Fuente: [5]

Ahora bien, la información que describe los recursos expuestos en la Web, se manejan a través de metadatos, que son esencialmente una descripción sobre los datos, materializado de manera eficiente mediante el uso de esquemas de representación de conocimiento como a) un conjunto de conceptos en un dominio y b) las relaciones entre estos conceptos; dichos esquemas reciben el nombre de ontologías (Figura 3), las cuales permiten organizar la información a través de vocabularios y taxonomías compartidas, permitiendo describir los recursos Web mediante la adición de descripciones sobre su semántica y sus relaciones:

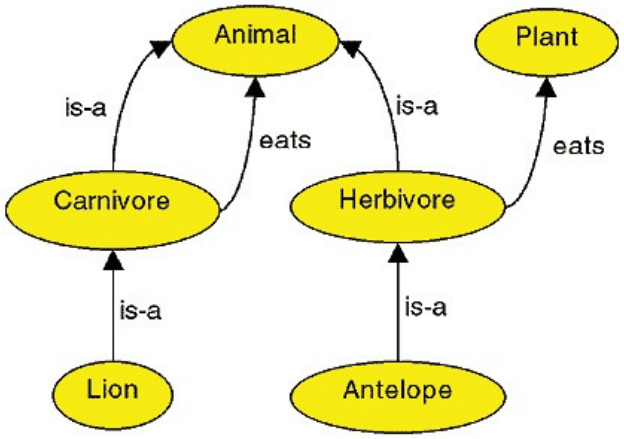

Fig. 3. Ejemplo básico de Ontología. Fuente: [12]

Al describir los recursos Web utilizando ontologías, se plantea como objetivo hacerlos legible para una máquina. Para llevar a cabo esta actividad, es necesario que cada (semántica) anotación corresponda a una pieza de información. Es importante en este proceso que la anotación de los recursos Web siga un estándar común [13], por ejemplo, el descrito en la Figura 4.

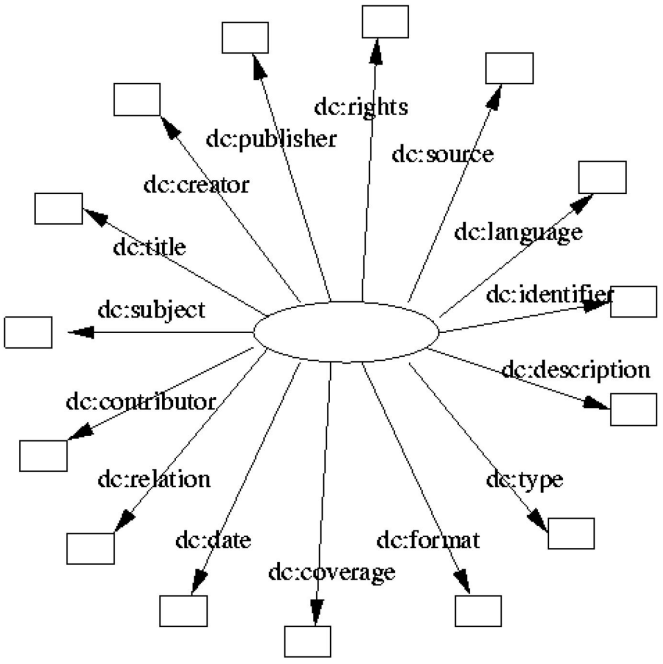

Fig. 4. Conjunto de Meta data Dublin Core. Fuente: [14]

Por otro lado, como lo describe [13], la publicación de estos datos puede hacerse en un formato propietario o en un formato abierto. Al publicarse en un formato abierto [15], la información se pone a disposición de todo el mundo para ser usada. El uso de URI y RDF para tales efectos es muy conveniente dado que los datos pueden estar unidos entre sí, creando un gran número de datos, que ofrece la capacidad de: buscar, combinar y explotar el conocimiento [16]. Los usuarios pueden incluso navegar entre las diferentes fuentes de datos, siguiendo los enlaces RDF, y navegar por un sitio Web potencialmente infinito de fuentes de datos conectadas

B. Principios Básicos de Linked Data

1. Esquema de Clasificación de 5 Estrellas

Tim Berners-Lee sugirió una clasificación de 5 estrellas [17], por medio de la cual los recursos van adquiriendo ciertas características, con el fin de llegar a un recurso plenamente vinculado, de acuerdo con los 4 principios básicos (figura 5). Esa jerarquía se compone de los siguientes 5 niveles.

- Publique la información en la Web, en cualquier formato, bajo un tipo de licencia de datos abiertos [18].
- Publique la información en forma de datos estructurados.
- Use formatos no propietarios.
- Use URIs para identificar las cosas, de manera que las personas puedan referenciar sus informaciones.
- Establezca la conexión entre sus datos y otros datos, con el fin de ofrecer un contexto ampliado para las informaciones.

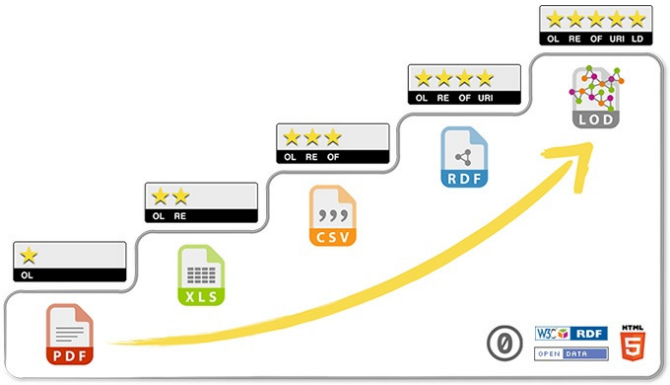

Figura 5. Esquema de 5 estrellas. Fuente: [17]

Niveles que presenta costos y beneficios incluyendo, por ejemplo, simplicidad de procedimientos, conversión de datos, formación de personal calificado, etc., todo eso con la intención de dirigirse hacia un conjunto de datos semánticamente mejor descritos, y conectados con otros datos.

C. Tecnologías de Linked Data

Para esbozar a muy alto nivel el contexto de LD, a continuación, se identifican un conjunto de tecnologías que interactúan, dentro de las cuales se destacan [19]:

a. URIs: Utilizadas para identificar cualquier cosa de interés en los datos, incluyendo las entidades de los datos, las clases o conceptos involucrados en los datos y las propiedades que están disponibles para describir.
b. RDF - Marco de Descripción de Recursos. Presenta la información como una serie de declaraciones simples. Cada declaración (tripleta) describe que algún sujeto tiene alguna propiedad con algún valor. El sujeto es normalmente identificado por un URI, así como la propiedad o predicado que describe. El valor de la propiedad y el objeto de la sentencia pueden tener un valor literal o puede ser otro URI.
c. Vocabularios Semánticos. Una parte fundamental del modelo de datos RDF son los vocabularios. RDF aumenta su capacidad de estructuración de la información si se combina con vocabularios específicos para la descripción semántica. Estos vocabularios RDF son definiciones de términos utilizados para efectuar los vínculos entre los diversos elementos de una descripción RDF.
d. SPARQL. Lenguaje de consulta para datos estructurados en la Web, específicamente datos accesibles en formato RDF o representables como tales. Por lo tanto es el lenguaje de consulta para los datos vinculados [20]. El propósito principal de SPARQL es proporcionar un lenguaje formal en el cual las preguntas significativas puedan ser formuladas [10].

D. Heterogeneidad e Interoperabilidad

Según como lo plantea [21], los recursos expuestos en la Web, se encuentran dispersos en una gran variedad de fuentes con diferente información, estructura y semántica, lo que ocasionan problemas de heterogeneidad, denotando así la necesidad de contar con una solución que permita la integración de los datos y el conocimiento embebido en los mismos, de manera eficiente. Es decir, se debe permitir a los usuarios acceder a los datos almacenados en fuentes de datos heterogéneas, presentando una única vista unificada de esos datos, de forma que el usuario no llegue a percibir esta heterogeneidad [5].

Para resolver el problema de la integración de las fuentes de información, se plantea la Interoperabilidad, entendida como "la capacidad de comunicarse, para ejecutar programas o para transferir datos entre varias unidades funcionales de una manera que requiera al usuario tener poco o nada de conocimiento de las características únicas de esas unidades (ISO 2382.1, 1993; ISO 19119, 2002)" [22]. Para alcanzar la denominada interoperabilidad, el modelo de información debe ser interoperable de forma:

- Sintáctica, que se refiere a la diferencia en el formato de datos, de modo que pueda ser procesada e interpretada en cualquiera de las fuentes de datos.
- Esquemática (estructural), que se refiere a las diferencias en el modelo de datos, en los esquemas, razón por la cual debe haber alguna forma de transformar de un esquema a otro.
- Semántica, que se refiere a las diferencias en la definición, en el significado que se pretende dar a los términos en contextos específicos. Este tipo de heterogeneidad se puede clasificar en: a) Heterogeneidad cognitiva: cuando no existe una base común en las definiciones para los fenómenos comunes en diferentes catálogos o bases de datos, y b) Heterogeneidad designativa: Hace referencia a fenómenos iguales semánticamente pero que son nombrados de una manera distinta. Aparecen a menudo en forma de homónimos, sinónimos o incluso polisemia.

Con el fin de abordar la heterogeneidad e interoperabilidad, se aboga por un enfoque de abajo hacia arriba, alentando a los desarrolladores a modelar de forma colaborativa datos, definir términos, vincular términos y conceptos a otros dataset heterogéneos, y utilizar bibliotecas de visualización genéricas y API para obtener más rápidamente aplicaciones útiles [24]. Para ello, es indispensable tener en cuenta los siguientes procesos [25]:

- Conversión: En primer lugar, los datos brutos (raw data) se limpian y se conservan a través de la representación basada en RDF. En segundo lugar, estos datasets convertidos utilizan URI desreferenciables, de manera que tanto los conjuntos de datos como sus ontologías pueden ser extendidos por terceros usuarios.

- Mejora: Se centra en extraer la semántica de valores literales en URIs significativas, y enlazar datasets asociando URIs mencionadas en diferentes datasets.

De tal forma que, vincular e integrar datos de formas novedosas ayuda a los consumidores a descubrir nuevos patrones y correlaciones y crear nuevos conocimientos. Los principios de datos vinculados y las tecnologías de Web Semántica facilitan la conexión de conjuntos de datos heterogéneos sin coordinación o planificación anticipada [24]. Para llevar a cabo esta actividad se busca reutilizar el vocabulario disponible, con el fin de hacer más eficientes los procesos de búsqueda, y acelerar el desarrollo de la Vinculación e integración. Esta actividad consta de las siguientes tareas [26]:

- Búsqueda de vocabularios adecuados. Existen algunos repositorios útiles para encontrar vocabularios disponibles, como por ejemplo Schema, SchemaCache, Swoogle y LOV. Para la elección de los vocabularios más adecuados se recomienda seguir las directrices propuestas en Open Data Commons Open Database License [27], tales como por ejemplo, tener cuidado con las declaraciones de los espacios de nombres de los vocabularios, que sirven para evitar que las definiciones de dos vocabularios se solapen [28].
- En caso de que no se encontrar ningún vocabulario adecuado, este se debe crear tratando de reutilizar gran parte de los posibles recursos existentes.
- Por último, si no se encuentra un vocabulario disponible ni los recursos para la construcción de la ontología, se debe crear la ontología desde cero.

Para muchos casos, los vocabularios ya se encuentran definidos y disponibles al público, y han sido ampliamente utilizados por muchas de las fuentes más importantes de datos enlazados disponibles en la actualidad. Entre ellos se destacan [29]:

- RDFS y OWL: contienen elementos que son bien conocidos y de utilidad al crear datos enlazados.
- FOAF: define los términos para describir a las personas, sus actividades y sus relaciones con otras personas. FOAF define un vocabulario de tipo RDF/XML para definir la información personal, como nombre, buzón y relaciones con amigos y otras propiedades.
- vCard. Este formato fue diseñado para la definición de las tarjetas de visita electrónicas. Es ampliamente utilizado por los clientes de correo electrónico y, a menudo utilizado para transmitir datos de contacto entre las personas u organizaciones. Dado que es un formato es muy popular, existe una codificación RDF que se puede utilizar al crear datos enlazados.
- BIBO. Ontología bibliográfica que ofrece los principales conceptos y propiedades para describir citas y referencias bibliográficas (es decir, libros, artículos, etc.).
- VIVO. Esta ontología ofrece un conjunto de tipos (clases) y relaciones (propiedades) para representar a los investigadores y todo el contexto en el que se desempeñan.

Estos vocabularios contienen tanto definiciones informales, en forma de documentación legible por humanos, como definiciones formales, en forma de reglas y restricciones que permiten detectar inconsistencias o deducir nuevos hechos a partir de otros dados (Figura 6).

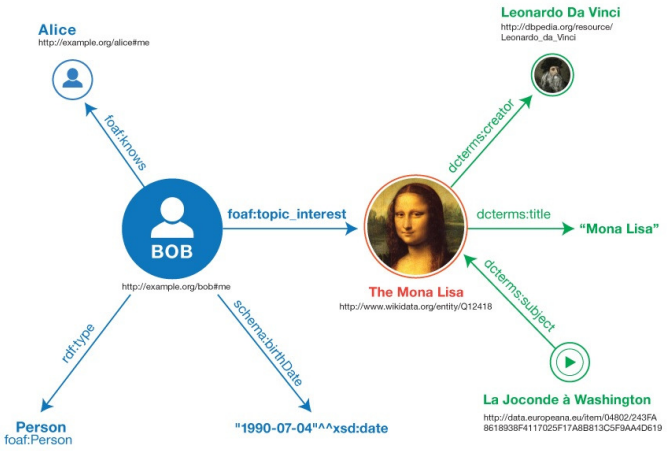

Figura 6. Grafo RDF usando Vocabularios. Fuente: [30].

Para ejemplificar el uso de vocabularios, una ontología puede definir clases para libros, pinturas y personas, o una propiedad "autor", y declarar formalmente que todos los recursos conectados a libros mediante la propiedad "autor" serán del tipo "persona". Puede igualmente definir un objeto de otra clase como una superclase de autor y pintura. Mediante un motor de inferencia que trabaje con los datos de una colección de libros y pinturas, y buscando por todos los objetos creados por una persona, se podrían recuperar todos ellos sin conocer a priori su tipología, una característica crucial para la integración de información [16].

E. Calidad de Datos

En primera instancia, y como lo plantea [31], la calidad de los datos es comúnmente concebida como una construcción multidimensional con una definición popular como la "aptitud para el uso". La calidad de los datos puede depender de varios factores como la precisión, puntualidad, integridad, relevancia, objetividad, credibilidad, comprensibilidad, consistencia, concisión, disponibilidad y verificabilidad. En términos de Web Semántica, existen diversos conceptos de calidad de los datos:

- Los metadatos semánticos, por ejemplo, es un concepto importante a considerar cuando se evalúa la calidad de los conjuntos de datos.
- Por otro lado, la noción de calidad de enlace es otro aspecto importante en los Datos Enlazados, donde se detecta automáticamente si un enlace es útil o no [5]. Asimismo, debe tenerse en cuenta que los datos y la información se utilizan indistintamente en la bibliografía.

En cuanto a los problemas de calidad de datos [31]:

- Bizer et al., define los problemas de calidad de los datos como la elección del diseño de sistemas de información basados en la Web que integran la información de diferentes proveedores. Adicionalmente clasifica las dimensiones de calidad de los datos en 3 categorías, de acuerdo al tipo de información que es usada como dimensión de calidad: (i) información basada en el contenido; (ii) información basada en el contexto; (iii) información basada en el Rating.
- Mendes et al., plantea que el problema de la calidad de los datos está relacionado con los valores que están en conflicto entre diferentes fuentes de datos como consecuencia de la diversidad de los datos.
- Flemming no proporciona una definición pero explica implícitamente los problemas en términos de diversidad de datos.
- Hogan et al., discuten sobre errores o ruidos o dificultades.
- Harth et al., discute acerca de los problemas de modelado que son propensos a las no explotaciones de esos datos de las aplicaciones.

Por lo tanto, el problema de la calidad de los datos se refiere a un conjunto de problemas que pueden afectar la potencialidad de las aplicaciones que utilizan los datos. Según [32], la calidad en los metadatos es un requisito previo para que los metadatos sean útiles.

La necesidad de que los metadatos estén adecuadamente definidos surge con el fin de facilitar que posteriormente se rellenen de forma correcta, para que se puedan explotar efectivamente. La calidad en los metadatos refleja el grado con el que los mismos realizan sus funciones esenciales de búsqueda, localización, uso, procedencia, autenticación y administración. Atendiendo a ello, [31] presentan una categorización de las dimensiones de calidad de los datos, con sus respectivas métricas (Figura 7).

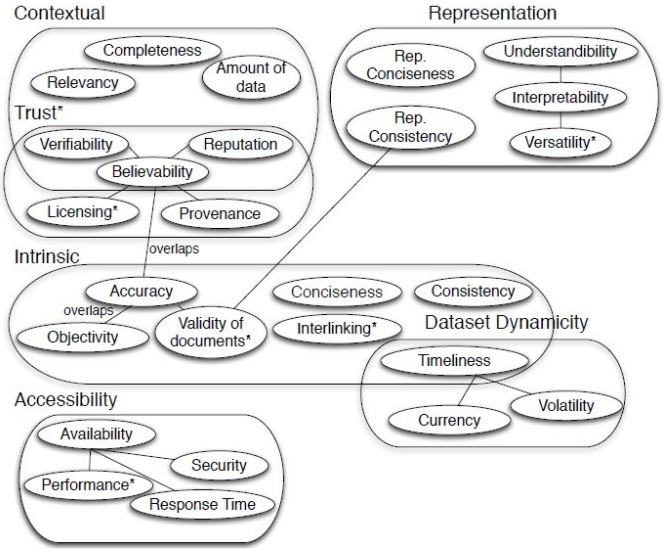

Figura 7. Dimensiones de Calidad de los Datos. Fuente: [31]

Sobre todo ello, es importante anotar que en el proceso de creación de los metadatos, los autores de los recursos deben tomar un papel fundamental. La creación descontrolada de metadatos provoca una pérdida de la interoperabilidad de metadatos entre repositorios digitales. Adicional a lo anterior, como se plantea en [33], uno de los principios fundamentales de datos enlazados es que las fuentes de datos contengan enlaces que apunten a información contenida en otras fuentes de datos, situación ante la cual muchas veces se presenta el problema de poseer enlaces rotos. Según Flemming y Hartig, los enlaces rotos pueden clasificarse en dos categorías:

- Enlaces estructuralmente rotos: se refieren a representaciones de recursos que no pueden recuperarse nunca más, y son fáciles de detectar de manera automática.
- Enlaces semánticamente rotos: se refieren a enlaces en los que la interpretación humana de lo que debe representar ese enlace difiere de la representación real. De la misma manera, estos autores definen tres tipos de causas por las cuales se crean enlaces rotos:
  - Tipo A: Originados porque se cambia la localización del recurso origen.
  - Tipo B: Originados porque se cambia la localización del recurso destino.
  - Tipo C: Originados porque se eliminan el recurso origen o el recurso destino.

Para mitigar este problema, el modelo de datos RDF proporciona los mecanismos necesarios para permitir conectar información de diferentes orígenes de datos en un único grafo global que posteriormente podrá ser procesado para obtener la información de las diferentes fuentes de datos. Además del entrelazado de los propios datos, también es necesario que los vocabularios utilizados por las distintas fuentes de datos sean relacionados con aquellos utilizados por otras. La situación ideal en el entrelazado de fuentes de datos es la utilización de

vocabularios comunes que se encuentren ampliamente aceptados en el dominio de datos correspondiente. Sin embargo, esto no es siempre posible ya que los proveedores de fuentes de datos enlazadas podrán necesitar la utilización de vocabularios mucho más específicos que los disponibles. Por ende, las diferentes fuentes de datos deberán relacionar los vocabularios específicos utilizados entre sí, permitiendo que se pueda navegar por dichos vocabularios y los datos correspondientes.

Por otro lado, la procedencia de los datos es uno de los aspectos principales a la hora de evaluar la autenticidad de un dataset, y por tanto, su confiabilidad, es la procedencia de los datos que lo componen. El W3C Provenance Incubator Group define la procedencia de un recurso de la Web de Datos como un registro que describe personas, entidades y procesos involucrados en la producción y lanzamiento, o que de otro modo hayan tenido influencia sobre dicho recurso. Por tanto, se considera la información de procedencia de un dataset como un registro de su historia, desde su creación, incluyendo información acerca de sus orígenes, y sus diferentes accesos y modificaciones.

IV. PLANTEAMIENTO DE LA DISCUSIÓN

Después de revisar la conceptualización previa, se identifica que la "Metadata" que describe un "Recurso", se basa en una "Taxonomía" que utiliza un "Vocabulario" controlado y reconocido. Adicionalmente, las taxonomías son particulares para los dominios que se vayan a trabajar, y estas deben analizar los elementos y características que propicien la interoperabilidad de vocabularios, ofreciendo recomendaciones para establecer y mantener mapeos entre ellos.

Atendiendo a dichas características, en la actualidad existen dominios que tienen taxonomías claramente definidas, y que en su mayoría manejan vocabularios controlados. Como lo plantean [34-37], dentro de estos dominios, se han definido conjuntos de datos prioritarios, como por ejemplo:

- Empresas.
- Seguridad y justicia.
- Recursos de la Tierra.
- Educación.
- Estadísticas nacionales.
- Mapas nacionales.
- Datos electorales a nivel nacional.
- Presupuestos nacionales.
- Finanzas y contratos.
- Geoespacial.
- Transparencia y Democracia.
- Sanidad.
- Ciencia e Investigación.
- Multilingüismo.
- Mapas nacionales.
- Energía y Medio Ambiente.
- Transporte e Infraestructura.

Dominios que puntean en la publicación de dataset que soportan la interoperabilidad de las fuentes que los contienen, como se evidencia en la Figura 8.

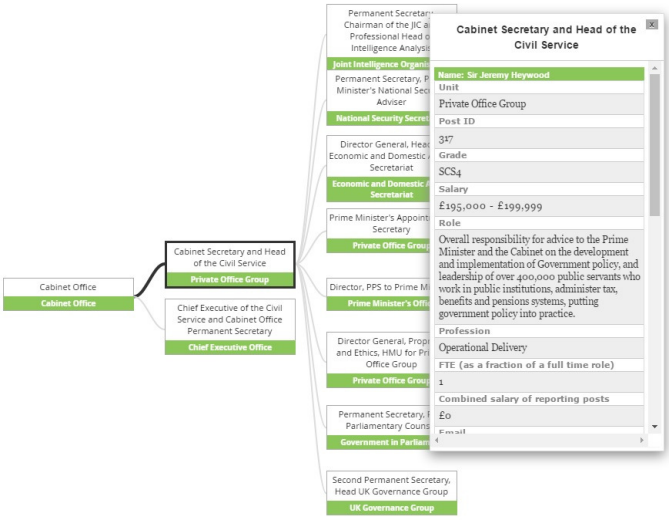

Figura 8. Ejemplo dominio de Datos Prioritarios. Fuente: [38].

Por otro lado, también se observa la cara opuesta de la moneda, donde se encuentran dominios con bajo nivel de madurez en sus procesos de estandarización, generación de vocabularios controlados, entre otros aspectos, que restringen o dificultan la creación de metadatos para la descripción de datos y documentos, como una de las principales técnicas para hacer los datos entendibles por agentes de software (máquinas) en la Web, y de esta forma hacer posible que sean compartidos de manera automática [29]. Por ende, la existencia de áreas de conocimiento para las cuales no se ha definido clara y totalmente una taxonomía, dificulta el expresar su metadata de forma adecuada, a través de un vocabulario que contribuya a la interoperabilidad.

La falta de madurez en este proceso se ve influenciada por diferentes factores como por ejemplo: la falta de voluntad política, problemas culturales, falta de leyes y normativas, falta de liderazgo, falta de personal idóneo, desconocimiento, desconfianza, entre otros [39]. Ahora bien, hay muchos caminos para expresar la metadata que describa un recurso, pero si esta metadata no responde a una taxonomía adecuada y con unos dimensiones de calidad adecuadas, llegaremos a alguna parte, siempre que caminemos lo suficiente, pero podría no ser el destino más apropiado, que asegure la interoperabilidad y la calidad entre los datos.

Además, es importante tener en cuenta que si el vocabulario no es adecuado o no existe, o en su defecto, la taxonomía para el dominio especifico no es clara o no existe, estos se deben construir desde cero, lo cual permitiría generar un lenguaje envolvente que se oriente a permear la interoperabilidad de los datos en la Web, y por ende permita a los usuarios poder

vincularlos de forma pertinente y oportuna. Un ejemplo clásico, donde se manifiestan los factores de heterogeneidad, interoperabilidad y calidad de datos se ve expuesto en [40], del cual se retoman los siguientes apartados:

- Los conjuntos de datos vienen a menudo sin nombre, sin descripción, sin propietario e inéditos. La gran cantidad de datos generados por diversas entidades en todo el mundo empeora el problema: ¿Cómo se puede encontrar una aguja en un montón de agujas si la aguja individual no puede describirse de una manera única?
- Los autores tiene la responsabilidad de asignar identificadores a sus datos, conectar los descriptores de metadatos a él, colocarlo en un repositorio confiable u otra ubicación predecible, y publicarlos
- Para abordar esta problemática, en primera instancia se debe dar nombre a los datos. Para algunos objetos digitales, como publicaciones científicas, el DOI (Identificador de objetos digitales) se ha convertido en estándar. En Europa, los investigadores suelen utilizar URI (identificadores uniformes de recursos).
- En segunda instancia, adicional a un identificador, se debe describir lo que hicieron los investigadores para reunir los datos: ¿Qué procedimientos de laboratorio utilizaron? ¿Qué máquinas tomaron las mediciones, secuencias determinadas o imágenes recogidas? ¿Qué significan ciertos campos de datos o siglas? Todo eso es opaco para el consumidor de los dataset, a menos que alguien haya descrito el dataset con metadatos.
- En tercera instancia se presenta el descubrimiento de los datos. Los buenos metadatos son un primer paso hacia el descubrimiento de datos. El siguiente es un índice y un motor de búsqueda que puede encontrar que los metadatos en respuesta a la consulta de un investigador.

En síntesis, en cuestiones de heterogeneidad, interoperabilidad y calidad de los datos, se debe considerar que:

a) Como lo manifiesta [31], LOD es la fusión de tres áreas de investigación diferentes:

- La Web Semántica, para generar conexiones semánticas entre los dataset,
- La World Wide Web, para poner disponibles los datos, preferiblemente bajo una licencia de acceso abierto, y
- Data Management, para manejar grandes cantidades de datos Heterogéneos y Distribuidos.

De allí que la construcción de una taxonomía envolvente, que permee todos los escenarios posibles del área de conocimiento específica, permitiría pasar del escenario procedimental (donde la taxonomía existe), al cubrimiento de un vacío de conocimiento, a través de la estructuración de dicha taxonomía, con su correspondiente identificación, construcción y uso de vocabulario, que asegura interoperabilidad entre las diferentes fuentes de datos.

b) La calidad de los datos es un asunto crucial, ya que proporciona la base para que los usuarios decidan si los datos a usar son seguros y confiables. A menudo, los usuarios deciden confiar o no en datos (Web) basados en el valor de dimensiones de calidad de metadatos, como exactitud, precisión y otros. La confianza es un fenómeno complejo que implica una actitud de un usuario hacia un tercero (que puede ser también un dato), seguido por una acción de confianza en un contexto especifico. Debido a su naturaleza subjetiva, las actitudes y acciones de confianza son desafiantes para estimar y predecir [41].

c) Evaluar la calidad de la información en la Web es un desafío por al menos dos razones: Primero, dado que la Web es una plataforma de publicación de datos descentralizada, en la que cualquier persona puede compartir casi cualquier cosa, no tiene mecanismos inherentes de control de calidad para asegurar que el contenido publicado sea válido, legítimo o incluso interesante [42, 43].

   En segundo lugar, al evaluar la fiabilidad de las páginas web, los usuarios tienden a basar sus juicios en criterios descriptivos como la presentación visual del sitio web en lugar de criterios normativos más sólidos como la reputación del autor, el proceso de revisión de la fuente, etc. Como resultado de ello, los usuarios de la Web son propensos a hacer evaluaciones incorrectas, particularmente a hacer juicios rápidos a gran escala. Por lo tanto, los usuarios de Web necesitan criterios de credibilidad y herramientas para ayudarles a evaluar la confiabilidad de la información de la Web con el fin de depositar confianza en ella [44].

d) Dado que la Web Semántica es grande, heterogénea, dinámica e incierta, la confianza será inevitablemente un problema. Dado su grado de distribución, contará con colecciones de agentes en dominios restringidos, lo cual implicaría que un agente sea responsable de reunir suficiente información para sus propios juicios de confianza, pero puede ser propenso al error, dependiendo de cuantas fuentes de información deba consultar y cuan confiables sean. Por lo cual, para que la confianza surja, los agentes deben ser flexibles y adaptarse a nuevos contextos [45,46, 47].

V. CONCLUSIONES

Como se presenta en [25], LD permite la interconexión de bases de datos, aportando descripciones de metadatos que en su integración permitirán inferir conocimiento. De ahí la importancia de la estandarización de la descripción de datasets y la conveniencia de la normalización de identificadores, lenguajes descriptivos, ontologías y vocabularios. De no evolucionar en ese camino, la no normalización puede impedir una interoperabilidad consistente de los datos.

Si los vocabularios, por ejemplo, carecen de la suficiente homogeneidad, las descripciones se podrán interconectar, pero no se podrá extraer nuevo conocimiento de ellas. Por ende, la interoperabilidad, entendida como la habilidad para interoperar o integrar diferentes conjuntos de datos, se convierte en un factor fundamental para asegurar la disponibilidad, acceso, reutilización, redistribución y participación universal en los datos.

Ahora bien, la Web de los Datos no es sino una versión para máquinas de la Web actual. Este modelo pretende que un software (llámese Agente, Aplicación, Asistente Personal, etc.), pueda interactuar con la Web sin necesidad de interpretar el lenguaje natural humano, y por tanto, beneficiarse de los datos allí publicados. Lo ideal es que en un futuro hubiese una versión en datos (LD Versión) de cualquier Web en formato documento (html, doc, etc.) que se orienta a humanos. Así, las máquinas pasarían a ser ciudadanos de primera categoría de la Web, cosa que ahora no son. ¿Pero, como conseguirlo? Esta interrogante se contesta a través del formateo de la información en tripletas RDF.

En consecuencia, hablar de heterogeneidad es intrínseco puesto que persé la información siempre será heterogénea: los datos del censo son diferentes a los datos meteorológicos, y más aún son diferentes a los datos producidos por un aparato conectado a la Internet de las Cosas (IOT). Los modelos de datos pueden ser distintos pero si el formato de datos es común (RDF+URIS desreferenciables+acceso a través de puntos SPARQL), la interoperabilidad es posible. Adicionalmente, considerando el desarrollo creciente que se ha observado en el ámbito del e-Learning [48], donde los recursos abiertos digitales adquieren mayor participación, exigiendo contenido con mejor contexto y de mejor calidad para los procesos de formación, la vinculación de recursos digitales abiertos cobra mayor importancia, y por ende, su interoperabilidad, con el fin de poder ser vinculados a través de la Web.

Acorde a los puntos de vista examinados anteriormente, se plantea como trabajos futuros: i) Profundizar en la identificación del lenguaje que circunscribe los recursos digitales abiertos, como área de conocimiento sobre la cual se lleva a cabo el proceso de investigación; ii) Identificación del vocabulario en Recursos Digitales Abiertos, que permita modelar los datos identificados en dicha área de conocimiento; iii) Estructurar y construir la ontología que describa la taxonomía de dichos recursos y permita expresar la metadata necesario para llevar a cabo su vinculación; iv) Diseñar un metamodelo para la vinculación de Recursos Digitales Abiertos, basado en LD, cuyos agentes tenga en consideración las dimensiones de cálida definidas en el proceso de investigación.

AGRADECIMIENTOS

Esta investigación se lleva a cabo en el marco de la formación doctoral en Ingeniería, en la Universidad Distrital Francisco José de Caldas. De igual forma, la temática planteada se configura como una línea de investigación de Grupo GIIRA.

REFERENCIAS

[1]. W3C. HTTP - Hypertext Transfer Protocol. [Online]. Available: https://www.w3.org/Protocols/.

[2]. W3C. Resource Description Framework (RDF). [Online]. Available: https://www.w3.org/RDF/.

[3]. M. Lamarca. XLL. s.f. [Online]. Available: http://www.hipertexto.info/documentos/xll.htm.

[4]. W3C. Naming and Addressing URIs, URLs. [Online]. Available: https://www.w3.org/Addressing/.

[5]. T. Berners-Lee. What is linked data? TED2009. TED.com. 2009. [Online]. Available: http://data.gov.uk/linked-data.

[6]. University of Mannheim. The Linked open Data Cloud Diagram. [Online]. Available: http://lod-cloud.net/.

[7]. W3C. Guía Breve de Linked Data. [Online]. Available: http://www.w3c.es/Divulgacion/GuiasBreves/LinkedData.

[8]. T3chFest. Creación de Agentes Inteligentes aplicando Tecnologías de la Web Semántica y Aprendizaje Automático. 2015. [Online]. Available: https://t3chfest.uc3m.es.

[9]. D. Sarango Romero. Publicación de datos universitarios observando los principios de Linked Data. Universidad Técnica Particular de la Loja. Ecuador. 2011. [Online]. Available: http://dspace.utpl.edu.ec/bitstream/123456789/1015/3/Tesis%20Sarango%20Romero%20Darwin%20Leonardo.pdf.

[10]. D. Wood et al. Linked data: Structured Data on the Web. Manning Publications, Shelter Island. ISBN 9781617290398. 2014.

[11]. Ontotext. What are Linked Data and Linked Open Data? [Online]. Available: http://ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/.

[12]. S. White. Better computational descriptions of science. Scientific Computing World. 2005. [Online]. Available: https://www.scientific-computing.com/feature/better-computational-descriptions-science.

[13]. N. Konstantinou, D. Spanos. Materializing the Web of Linked Data. Springer. National Technical University of Athens. ISBN 978-3-319-16073-3. Athens, Greece. 2015.

[14]. J. Herrera-Cubides et al. Standardization Initiatives in the Production of Virtual Learning Objects. JISTEM - Journal of Information Systems and Technology Management. Vol. 11, No. 3, Sept/Dec. 2014.

[15]. W3C. Linked Open Data. [Online]. Available: https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

[16]. M. Doerr et al. El Modelo de Datos de Europeana (EDM). 2010. [Online]. Available: http://conference.ifla.org/past-wlic/2010/149-doerr-es.pdf.

[17]. 5StarData. 5 Stars Open Data. s.f. [Online]. Available: http://5stardata.info/en/.

[18]. BNCC. Biblioteca del Congreso Nacional de Chile. Linked Open Data: Qué es?. s.f. [Online]. Available: http://datos.bcn.cl/es/informacion/que-es.

[19]. R. Ávila Alonso. Aplicación de los principios Linked Open Data a la lista de encabezamientos de materia de la Biblioteca de la Universidad Politécnica de Madrid. Universidad Carlos III de Madrid. 2014. [Online]. Available: https://core.ac.uk/download/pdf/29406301.pdf.

[20]. A. Graves. Creando Aplicaciones Basadas en Linked Data. 2012. [Online]. Available: http://manzanamecanica.org/2012/01/tutorial_creando_aplicaciones_basadas_en_linked_data_parte_13.html.

[21]. S. Speicher et al. Linked Data Platform 1.0. W3C Recommendation. Vol 1. 2015.

[22]. F. Flores. Integración y publicación como Open Linked Data, de información geográfica catastral a través de ontologías bajo el contexto de la web semántica. Universidad Nacional de Colombia. Bogotá, Colombia. 2015. [Online]. Available: http://www.bdigital.unal.edu.co/51274/1/52875927.2015.pdf.

[23]. R. Míguez Pérez et al. Linked Data como herramienta en el ámbito de la nutrición. Departamento de Ingeniería Telemática. Universidad de Vigo. España. Nutrición Hospitalaria. Vol.27 no.2 Madrid. ISSN 1699-5198. 2012. [Online]. Available: http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-16112012000200001.

[24]. D. DiFranzo et al. TWC LOGD: A Portal for Linking Open Government Data. SWC 2010. 2010. [Online]. Available: http://challenge.semanticweb.org/submissions/swc2010_submission_16.pdf.

[25]. L. Ding et al. Data-Gov Wiki: Towards Linking Government Data. s.f. [Online]. Available: https://pdfs.semanticscholar.org/22a9/b850e3aa6eaa67744b65fd9ea135e4eae3ce.pdf.

[26]. D. Wood. Linking Government Data. Springer New York Dordrecht Heidelberg London. Library of Congress. E-ISBN 978-1-4614-1767-5. 2011.

[27]. ODbL. Open Data Commons Open Database License (ODbL). [Online]. Available: https://opendatacommons.org/licenses/odbl/.

[28]. R. Saquete. El Impredecible Futuro de la Web Semántica. 2013. [Online]. Available: http://www.humanlevel.com/articulos/desarrollo-web/el-futuro-de-la-web-semantica.html.

[29]. G. Bustos. Prototipo de un sistema de integración de recursos científicos, diseñado para su funcionamiento en el espacio de los datos abiertos enlazados para mejorar la colaboración, la eficiencia y promover la innovación en Colombia. Universidad Nacional de Colombia. Bogotá, Colombia. 2015. [Online]. Available: http://www.bdigital.unal.edu.co/50580/1/MSc_2702295_V50.pdf.

[30]. F. Manola et al. RDF 1.1 Primer. W3C Working Group Note 24. 2014. [Online]. Available: https://www.w3.org/TR/rdf11-primer/.

[31]. A. Zaveri et al. Quality Assessment Methodologies for Linked Open Data. A Systematic Literature Review and Conceptual Framework. IOS Press. 2012. [Online]. Available: http://www.semantic-web-journal.net/content/quality-assessment-methodologies-linked-open-data.

[32]. D. Pons et al. La estandarización para la Calidad en los Metadatos de Recursos Educativos Virtuales. Universidad de Alcalá. [Online]. Available: http://www.esvial.org/wp-content/files/estandarizacionMetadatosPonsHileraPages.pdf.

[33]. BizkaiLab. Estado del arte en confianza y calidad de fuentes de datos enlazadas. Proyecto BIDEI. 2011. [Online]. Available: http://www.bizkailab.deusto.es/wp-content/uploads/2012/04/5761.pdf.

[34]. red.es. Tendencias actuales en iniciativas Open Data. Ministerio de Industria, Energía y Turismo. Gobierno de España. 2014. [Online]. Available: http://datos.gob.es/sites/default/files/bestpractices_opendata_sep2014_1_1.pdf.

[35]. EU. EU implementation of the G8 Open Data Charter. G8 Open Data Charter. 2013. [Online]. Available: https://ec.europa.eu/digital-single-market/news/eu-implementation-g8-open-data-charter.

[36]. European Commission. Report on high-value datasets from EU institutions SC17DI06692. European Commission. 2014. [Online]. Available: https://joinup.ec.europa.eu.

[37]. EU. European Union Open Data Portal. s.f. [Online]. Available: https://data.europa.eu/euodp/en/data.

[38]. ODUG. Open Data Request Map. Open Data User Group (ODUG). [Online]. Available: https://data.gov.uk/organogram/cabinet-office.

[39]. OD-MM. MMOD Modelo de Madurez open Data. [Online]. Available: https://sites.google.com/a/ouiiohe.org/datos-abiertos/el-proyecto#TOC-Resultados-gen-ricos.

[40]. K. Miller. Data's Identity Crisis: The Struggle to Name It, Describe It, Find It, and Publish It. Biomedical Computation Review. 2016. [Online]. Available: http://biomedicalcomputationreview.org/sites/default/files/dataidentity_bcr-spring-2016-web.pdf.

[41]. D. Ceolin et al. Linking Trust to Data Quality. 2015. [Online]. Available: http://www.few.vu.nl/~dceolin/method2015.pdf.

[42]. J. Herrera-Cubides, P. Gaona-García, K. Gordillo-Orjuela. A View of the Web of Data. Case Study: Use

of Services CKAN. Ingeniería, v. 22, n. 1, p. 111-124. 2017. ISSN 2344-8393. [Online]. Available at: https://revistas.udistrital.edu.co/ojs/index.php/reving/article/view/10542/12408.

[43]. J. Herrera-Cubides, P. Gaona-García and S. Sánchez-Alonso. The web of data: Past, present and ¿future? XI Latin American Conference on Learning Objects and Technology (LACLO), San Carlos, 2016, pp. 1-8. DOI: 10.1109/LACLO.2016.7751802. [Online]. Available: https://ieeexplore.ieee.org/document/7751802/.

[44]. J. Pattanaphanchai et al. Trustworthiness Criteria for supporting users to assess the Credibility of Web Information. 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2488132.

[45]. K. OHara et al. Trust Strategies for the Semantic Web. s.f. [Online]. Available: https://pdfs.semanticscholar.org/c771/5a87e1dc757abbcd62f4afc7c70c88b8d9e2.pdf.

[46]. J. Herrera-Cubides et al. A Fuzzy Logic System to Evaluate Levels of Trust on Linked Open Data Resources. Revista Facultad de Ingeniería, n. 86, p. 40-53. 2018. ISSN 2422-2844. [Online]. Available at: http://aprendeenlinea.udea.edu.co/revistas/index.php/ingenieria/article/view/328937. DOI: http://dx.doi.org/10.17533/udea.redin.n86a06.

[47]. E. Arias-Caracas, D. Mendoza-López, P. Gaona-García, J. Herrera-Cubides, C. Montenegro-Marín. Evaluation of the Linked Open Data Quality Based on a Fuzzy Logic Model. 2018. Artificial Intelligence Applications and Innovations. AIAI 2018. IFIP Advances in Information and Communication Technology, vol 519. Springer. [Online]. Available at: https://link.springer.com/chapter/10.1007/978-3-319-92007-8_47

[48]. C. Pappas. The Top eLearning Statistics and Facts for 2015 you need To Know. E-Learning Industry. 2015. [Online]. Available: http://elearningindustry.com/elearning-statistics-and-facts-for-2015.