Revista Logos Ciencia & Tecnología, 16(2), 171-185 | Mayo-agosto, 2024
https://revistalogos.policia.edu.co:8443/index.php/rlct

Artículo de reflexión

Revista Logos Ciencia & Tecnología
ISSN 2422-4200 (en línea)

El papel de los datos abiertos vinculados (LOD): una mirada desde las fuentes de datos inteligentes (OSINT)

The role of linked open data (LOD): A perspective based on open-source intelligence (OSINT)

O papel dos dados abertos vinculados (LOD): um olhar sobre a inteligência das fontes abertas (OSINT)

Álvaro Varón-Capera [a,*] | Paulo Alonso Gaona-García [b] | Carlos Enrique Montenegro-Marín [c]

a. https://orcid.org/0009-0005-4134-7304 Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia
b. https://orcid.org/0000-0002-8758-1412 Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia
c. https://orcid.org/0000-0002-3608-7158 Universidad Distrital Francisco José de Caldas, Bogotá D.C., Colombia

- Fecha de recepción: 2024-02-26
- Fecha concepto de evaluación: 2024-05-13
- Fecha de aprobación: 2024-05-30
https://doi.org/10.22335/rlct.v16i2.1944

Para citar este artículo/To reference this article/Para citar este artigo: Varón-Capera A., Gaona-García, P.A. & Montenegro-Marín, C. E. (2024). El papel de los datos abiertos vinculados (LOD): una mirada desde las fuentes de datos inteligentes (OSINT). *Revista Logos Ciencia & Tecnología, 16*(2), 171-185. https://doi.org/10.22335/rlct.v16i2.1944

RESUMEN

Cada día se generan millones de fuentes de datos abiertos disponibles en la web, que mediante diversos procesos pueden ser viables como fuentes de datos inteligentes para la toma de decisiones y actividades relacionadas con su consumo. En el presente artículo se plantea un estudio basado en una metodología estadística descriptiva, para realizar un análisis del crecimiento de datos abiertos enlazados, siguiendo los principios de "vinculación de datos abiertos" (LOD) y su potencial enmarcado dentro de las "fuentes de datos inteligentes" (OSINT). El estudio pretende analizar las estructuras que posibiliten determinar una serie de lineamientos y permitan verificar la validez y vinculación de fuentes inteligentes (OSINT) mediante principios basados en LOD. Finalmente, se *presentan algunas recomendaciones y acciones que permitan determinar a los actores principales de las distintas fuentes de datos para agregar cadencia al proceso de liberación de recursos LOD.*

Palabras clave: datos abiertos vinculados, web semántica, RDF, datos abiertos inteligentes, OSINT.

ABSTRACT

Everyday millions of open data sources available on the Web are generated, which through several and different processes can be viable as intelligent data sources for decision making and activities related to their consumption. This article proposes a study based on a descriptive statistical methodology, to carry out an analysis of the growth of linked open data, under the principles of Linking Open Data (LOD) and its potential through intelligent data sources (OSINT). The study aims to analyse the structures that allow a series of guidelines in order to verify the validity and linkage of OSINT under principles based on LOD. Finally, some recommendations

*Autor de correspondencia. Correo electrónico: avaronc@udistrital.edu.co

and actions are presented to determine the main actors of the different data sources to add cadence to the LOD resource release process.

Keywords: Linked open data (LOD), semantic web, open-source intelligence (OSINT).

RESUMO

A cada dia são gerados milhões de fontes de dados abertas disponíveis na web, que, através de diversos processos, podem ser viáveis como fontes de dados inteligentes para a tomada de decisões e para atividades relacionadas com o seu consumo. No presente artigo, é apresentado estudo baseado numa metodologia estatística descritiva, para realizar uma análise do crescimento dos dados abertos vinculados (LOD), seguindo seus princípios e seu potencial marcado dentro da inteligência das fontes abertas (OSINT). O estudo pretende analisar as estruturas que podem determinar uma série de linhas e permitir verificar a validade e a ligação das fontes inteligentes (OSINT) por meio de princípios baseados em LOD. Por fim, são apresentadas algumas recomendações e ações que permitem determinar os atores principais das fontes de dados distintas para dar continuidade ao processo de libertação de recursos LOD.

Palavras-chave: dados abertos vinculados (LOD), web semântica, RDF, inteligência das fontes abertas (OSINT).

Introducción

La búsqueda de información en internet ha representado grandes desafíos asociados a la efectividad de los resultados obtenidos, de acuerdo con los criterios de búsqueda de cada persona en un área de conocimiento. Este gran reto se debe, en parte, a la información que no está bajo un esquema estructurado donde se pueda clasificar debido a la gran variedad de formatos (Bizer et al., 2011), lo que lleva a resultados con múltiples salidas que no son acordes con la búsqueda efectuada.

La iniciativa propuesta por Tim Berners Lee (Bizer & Berners-Lee, 2008), asociada a la vinculación de datos mediante *linked data* (LD), promete resolver los problemas relacionados con el análisis y la interoperabilidad de los datos vinculados a recursos que se encuentren en internet. Así, la web semántica se concibe como una red extendida con mayor significado, donde cualquier usuario puede hallar la información que necesita de manera práctica y efectiva, porque la información está mejor definida, con mayor significado y con un grado de semántica mayor (Bizer et al., 2011). Con base en ello, se pueden obtener soluciones a problemas habituales en la búsqueda de información gracias a la utilización de una infraestructura común, mediante la cual es posible compartir, procesar y transferir información de forma sencilla, a partir de una serie de principios de vinculación de datos.

Es así como actualmente podemos identificar una gran variedad de servicios ofrecidos en la web, lo que ha generado una evolución y crecimiento de exposición de datos digitales (Pune, 2020). Estos datos pueden ser accesibles por API (*application programming interface*) o diferentes servicios, aplicaciones, entre otros, lo cual representa una fuente de recursos ideal para ser utilizados posteriormente para la generación de conocimiento, toma de decisiones y otras actividades enmarcadas dentro de las fuentes de datos inteligentes (OSINT, por sus siglas en inglés, *open source inteligence*). La motivación de este estudio es llevar a cabo un análisis del crecimiento de datos abiertos vinculados (LOD, por sus siglas en inglés *linked open data*) y su potencial para ser reutilizados como fuentes de datos inteligentes (OSINT).

Marco teórico

Datos abiertos

De acuerdo con Bikakis (2013), los datos abiertos se pueden definir como datos representados y liberados digitalmente, bajo una serie de características que permiten su uso, reutilización y redistribución de forma libre en cualquier escenario para cualquier fin.

Web semántica

Información alojada en la web con una sintaxis definida y con la información semántica nece-

saria para que sus datos sean comprendidos y procesados por computadoras. Así se agrega un significado a los recursos web de manera autónoma.

Datos abiertos vinculados

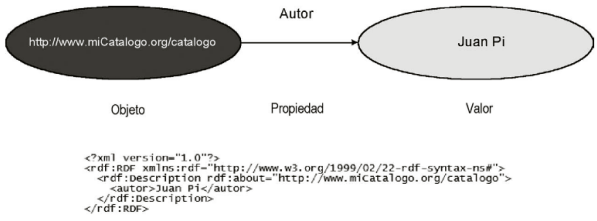
Serie de principios que permite la capacidad de enlazar conjuntos de datos con otros asociados, para cumplir la promesa de valor de la vinculación de datos. Un conjunto de datos debe utilizar URI para identificar las entidades que lo componen mediante el protocolo HTTP. Proveerse mediante la estructura RDF (*resource description framework*) que implementa una sintaxis estándar, enlazar estos datos con otros RDF que enriquezcan y describan la información.

RDF (*resource description framework*)

Es un marco de herramientas que permite intercambiar información en la web mediante un lenguaje formal que describe la información estructurada. Asimismo, permite a los recursos de la web analizar y procesar la información contenida en un conjunto de datos. RDF agrupa los recursos, lo que los convierte en grafos dirigidos, que en su enlace con otros objetos forman estructuras jerárquicas de árboles, convirtiéndolo en un lenguaje de descripción para las redes, y sus recursos deben tener una URI y estar estructurados en esquemas XML (véase Figura 1).

Figura 1

Representación gráfica de un grafo RDF y su estructura en esquema XML



Fuente: Peis et al. (2009).

OSINT

Como lo describe Carcaño (2018), las fuentes abiertas no solo provienen de internet, ni necesariamente tienen un soporte tecnológico. Las fuentes abiertas son de carácter público e independientes, ya sea que su contenido se comercialice o sea gratuito. Este tipo de fuentes, siempre disponibles para el público, pueden consistir en documentos de cualquier contenido, en cualquier soporte (papel, fotográfico, magnético…), a través de cualquier medio de transmisión (sonoro, audiovisual, impreso…), o modo de acceso, ya sea digital o no. Algunos ejemplos de este tipo de fuentes incluyen una agencia de noticias, un diccionario, un blog, un documento de tesis, una conferencia de prensa, un artículo científico, un canal RSS, una jurisprudencia, las emisiones de radio y televisión, entre otros.

Existen diversas definiciones de OSINT; sin embargo, en concordancia con los objetivos del presente artículo, se tomará como punto de partida la definición propuesta por Pastor-Galindo et al. (2020), quienes la definen como el proceso de "recopilar, procesar y correlacionar información pública de fuentes de datos abiertos como medios de comunicación, redes sociales, foros y blogs, datos del gobierno público, publicaciones o datos comerciales", lo anterior con el propósito de transformarlas a fuentes de datos inteligentes.

Contexto de LOD

Las iniciativas establecidas por las organizaciones para vincular datos abiertos (TLODC, por sus siglas en inglés *the linked open data*) (TLODC, 2024) son de gran importancia, dado que permiten llevar a cabo una categorización de las diferentes fuentes de datos abiertos vinculados y que posibilitan identificar recursos relacionados y compartidos de manera abierta por colaboradores y organizaciones en internet. Esta vinculación de recursos es un importante avance hacia el modelo ideal de la web semántica. Es así como el estilo de web clásico, que permitió interconectar varios sitios mediante hipervínculos con el objetivo de lograr una navegación para una búsqueda en internet, recibió una leve transformación a partir de iniciativas definidas en *linked data*, propuesta establecida por el mismo creador de la web, Tim Berners Lee (Bizer et al., 2011). *Linked data* es una forma de utilizar la red creando *links* entre datos de diferentes fuentes

de información. Técnicamente se refiere a los datos publicados en la red y que de alguna manera están explícitamente definidos; estos están conectados a otros repositorios de datos y, por consiguiente, pueden ser vinculados desde esos repositorios.

Las consultas que se efectúan actualmente en la red toman esos hipervínculos y los relaciona, arrojando muchas veces el resultado esperado, de allí que lleve tanto tiempo implementado. En muchas ocasiones, el problema con estos hipervínculos consiste en que no siempre arrojan resultados de búsqueda en el contexto de los criterios establecidos para lo mismo. Es así como *linked data* ha sido una de las iniciativas en las que se viene trabajando desde hace más de una década, con el propósito de establecer nuevos paradigmas para compartir recursos y fuentes de datos sobre internet. Estas estrategias, desarrolladas a partir de iniciativas definidas por *linked data*, son claves para refinar las búsquedas mediante la apertura de repositorios de datos, así como otras herramientas. Algunos de estos repositorios han sido acogidos por la iniciativa de datos abiertos a través del proyecto LOD Cloud (LOD2, 2024); proyecto que ha elaborado históricamente un diagrama que permite representar el número *datasets* publicados por diversos colaboradores y algunas organizaciones y compartidos bajo especificaciones *linked data*. La implementación de estas iniciativas ha permitido acercarnos a la idea que se ha planteado de utilizar una web semántica, que consiste en enriquecer recursos compartidos a través de metadatos, y relacionarlos con recursos que tengan algún grado de afinidad. Esto nos ha permitido establecer una comunidad entre colaboradores y organizadores para desarrollar modelos de comunicación que faciliten vincular una gran cantidad de fuentes de datos abiertos con esta iniciativa.

La siguiente sección brindará un panorama del crecimiento que ha tenido esta iniciativa, con el propósito de identificar estrategias que puedan enmarcarse dentro de las fuentes de datos inteligentes (OSINT) para facilitar su lectura y análisis.

Métodos utilizados para recopilación y organización de datos

El objeto de análisis para los conjuntos de datos y su relación con OSINT es la plataforma LOD Cloud, que, desde los distintos orígenes de datos abiertos vinculados, permite el consumo de sus datos mediante consultas por API y acceso bruto mediante URL a los recursos en información plana y estructurada JSON. A partir de su descarga y utilizando herramientas de apoyo para consumo de API como Postman, se almacenan los recursos en repositorios locales y utilizando la herramienta LODCloudDraw (LOD Cloud Raw, 2024) se verifica la consistencia de la información descargada.

Como estrategia, se recopilaron y validaron conjuntos de datos LOD bajo todos aquellos recursos que han sido vinculados en la plataforma LOD Cloud. Se organizaron y prepararon para identificar el crecimiento de los dominios o subnubes que forman la plataforma, del mismo modo su comportamiento dinámico a lo largo del tiempo. Posteriormente, se pretende identificar las bondades de OSINT, sus diferentes fuentes de datos, así como aquellas herramientas abiertas que se utilizan para la extracción de recursos. La Figura 2 presenta un resumen de la metodología utilizada para el análisis.

Figura 2

Resumen contextual de la metodología utilizada

Para su análisis, se aplicó un método cuantitativo descriptivo, el cual permite al lector identificar los criterios de evaluación con las muestras representativas definidas, jerarquizadas y relacionadas en las generalizaciones,

apoyada de gráficos para la interpretación e identificación de tendencias. Finalmente, se expondrán los resultados y análisis donde estas dos grandes iniciativas podrán ser utilizadas en conjunto para potenciar la extracción de datos e información para diversas áreas de conocimiento, donde se identifica que, de acuerdo con los resultados obtenidos, desde su creación hasta finales del 2018 existió un crecimiento exponencial en la plataforma LOD Cloud, permitiendo la habilitación de casi una decena de dominios o subnubes; pero desde el 2019 al 2023 se presenta un decremento en la actividad de las fuentes de datos, afectando la actualización de los repositorios y la alimentación de distintos conjuntos de datos.

Con los conjuntos de datos, también es posible identificar los dominios y subdominios que componen LOD Cloud, permitiendo llevar a cabo un análisis de su crecimiento a lo largo del tiempo, apoyado en gráficos liberados por la misma plataforma, que pueden ser contrastados con la herramienta LODCloudDraw.

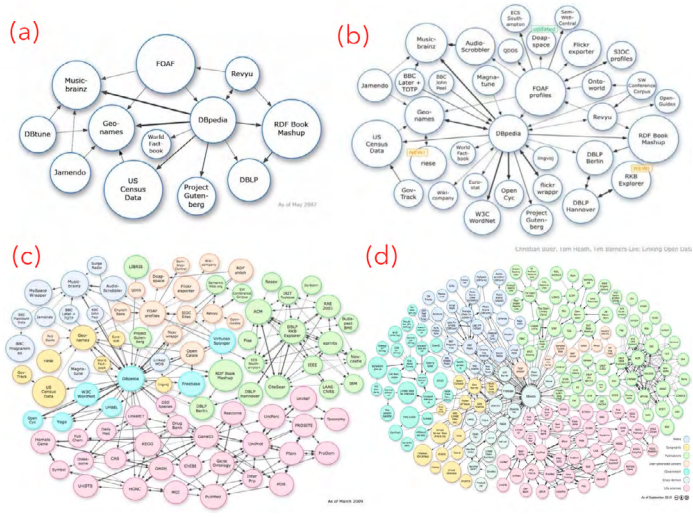
Análisis y visualización

Desde sus inicios en el 2007 (véase Figura 3a), LOD Cloud contaba con 12 *datasets* de información vinculada, cerca de 5 billones de tripletas RDF y 120000 *links* RDF (Bizer et al., 2011), como se muestra en la Figura 3 tomada de the linked open data cloud TLODC (2024).

Para el 2008 (véase Figura 3b), la iniciativa que se potencializaba, logró registrar 45 *datasets*, 2 billones de tripletas RDF y 3 millones de *links* RDF. Durante estos dos primeros años, no existía un grafo tan complejo como para colorear y exponer los nodos según el campo de conocimiento al que pertenecieran. Pero a partir del 2009, la red de conocimiento vinculada empieza a complejizarse y crecer al punto de necesitar clasificar los nodos por áreas de conocimiento. Para el 2009 (véase Figura 3c), de acuerdo con TLODC (2024), LOD Cloud contaba con 93 *datasets*, doblando la suma del año anterior, con 4.7 billones de tripletas RDF y 142 millones de *links* RDF. Estas cifras representan un significativo aumento desde el inicio de LOD Cloud, que en tan solo dos años ha logrado aumentar exponencialmente su contenido. Sin embargo, cerca de la cuarta parte de la población pertenecía a la comunidad de internet, que llevaba a cabo búsquedas diarias, para lo cual 4.7 billones de tripletas relacionadas con 93 organizaciones, no representaba una amplia y completa colección de datos globales. En el diagrama tomado de TLODC (2024a), este esfuerzo mancomunado de crecimiento lleva a que en el 2010 (véase Figura 3d) LOD Cloud tenga 203 *datasets*, algo más de 26 billones de tripletas RDF y cerca de 503 millones de *links* RDF. Por si los números no dan una idea de la cantidad de información condensada en LOD Cloud para este año, en el diagrama se puede ver cómo las relaciones se ven más complejas, incrementando las áreas de conocimiento vinculadas, y se expande el número de *datasets*.

Figura 3

Crecimiento LOD Cloud 2007-2010



Fuente: TLODC (2024).

Para el 2011, aunque se mantienen las siete áreas de conocimiento del año anterior, existe un amplio crecimiento en el contenido relacionado con algunas. En la Tabla 1 se relaciona la cantidad de *datasets p*or dominio o área de conocimiento*, links RDF* y tripletas asociadas. Ariba la tabla del 2010 y abajo la tabla del 2011.

A partir de este contexto preliminar 2007-2010, el aumento más dramático de *datasets* sucede en las áreas de gobierno y contenido generado por el usuario, con un aumento de 96% y 186%, respectivamente. Mientras que en áreas como ciencias de la vida aumenta el número de *datasets* en uno. Esta relación de incremento de *datasets* se puede observar en la Figura 4.

Tabla 1

Relación de datasets, tripletas 2010 y 2011

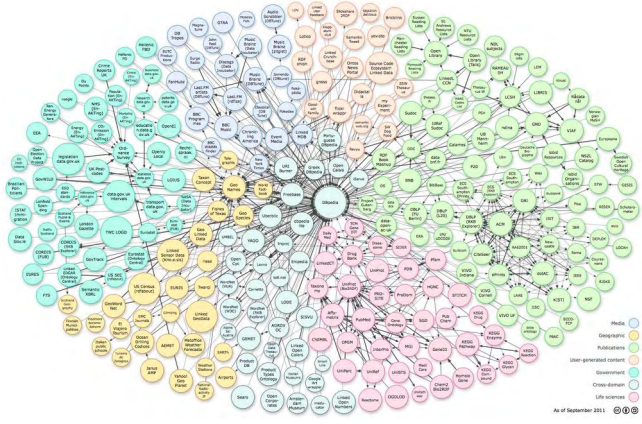
Domain	Number of databases	Triples	%	(Out)-Links	%
Media	26	2,453,898,811	9.11 %	50,374,304	12.74 %
Geographic	16	5,904,980.833	21.93 %	16,589,086	4.19 %
Government	25	11,613,525,437	43.12 %	17,658,869	446 %
Publications	67	2,237,435,732	8.31 %	77,951,898	19.71 %
Cross-domain	40	1,99,085,950	7.42 %	29,105,638	7.36 %
Life sciences	42	2,664,119,184	9.89 %	200,417,873	50.67 %
User-generated Content	7	57,463,756	0.21 %	3,402,228	0.86 %
		26,930509,703		395,499,896	

Domain	Number of databases	Triples	%	(Out)-Links	%
Media	25	1,841,852,061	5.82 %	50,440,705	10.01 %
Geographic	31	6,145,532,484	19.43 %	35,812,238	7.11 %
Government	49	13,315,009,400	42.09 %	19,343,519	3.84 %
Publications	87	2,950,720,693	9.33 %	139,925,218	27.76 %
Cross-domain	41	4,184,635,715	13.23 %	63,183,065	12.54 %
Life sciences	41	3,036,336,004	9.60 %	191,844,090	38.06 %
User-generated Content	30	134,127,413	0.42 %	3,449,143	0.68 %
	295	31,634,213,770		503,998,829	

Fuente: LOD2 (2024).

Figura 4

Crecimiento 5 LOD Cloud 2011



Fuente: TLODC (2024a).

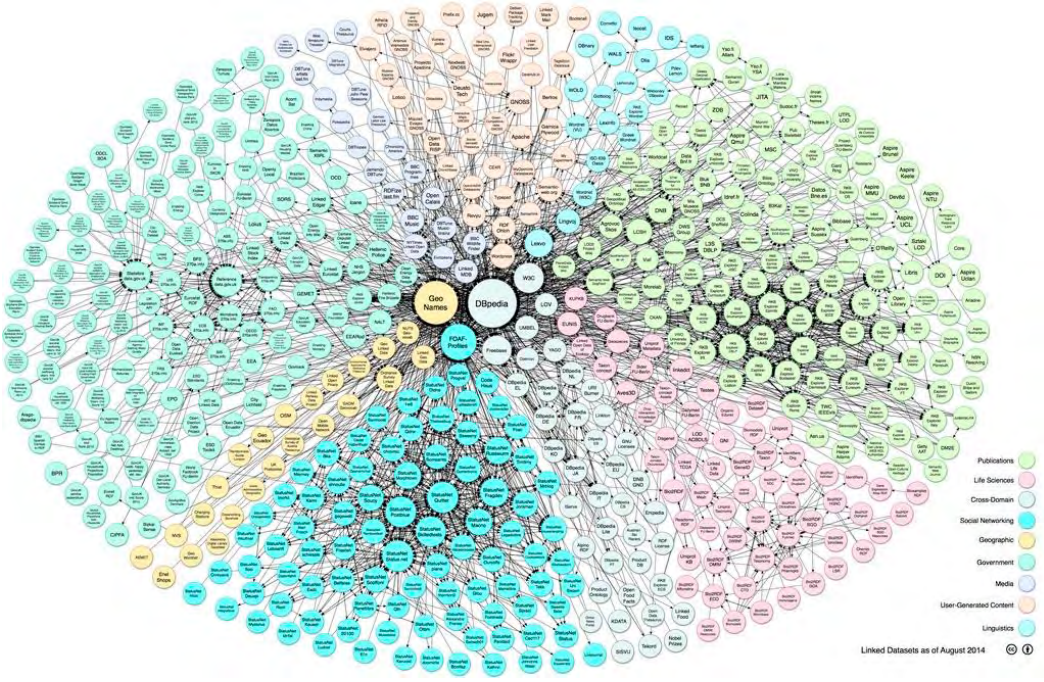
Conscientes de la cantidad de información presente sin vincular ni catalogar en la web, los adeptos a LOD Cloud, su fundador y todos los involucrados, continuaron trabajando por vincular la información en la web, garantizando la confiabilidad e integridad de la información. Uno de los campos inexplorados y muy importante de enlazar eran (y siguen siendo) los archivos gubernamentales, en donde la Berners-Lee desempeñó un papel importante junto a Open Knowledge Foundation (OKFN), dado que en el 2012 convencen y apoyan al Gobierno inglés para sumarse a la iniciativa (LOD2, 2024).

Para el 2014, el cambio en el número de *datasets* es muy alto. Se pasa de 295 *datasets* en el 2011 a 592 en el 2014 (Schmachtenberg et al., 2014). El 2012 y 2013 están débilmente documentados, y no fue posible establecer una cifra oficial de la cantidad de datos existentes en LOD Cloud. La Figura 5 permite ver cómo aumentaron también las áreas de conocimiento con contenidos LOD, entre las novedades se encuentran las redes sociales y lingüística.

Durante el 2017, LOD Cloud experimentó un gran salto, ya que, durante los tres años anteriores, desde el 2014, no se había registrado un crecimiento significativo; se logró un hito importante con la publicación de más de 1000 conjuntos de datos en LOD Cloud. Al finalizar el año, se contabilizaron 1163 *datasets*, lo que evidencia la consolidación de la comunidad de LOD Cloud. Además, se llevaron a cabo eventos y talleres para fortalecer esta comunidad y se integraron tecnologías como RDF (Wylot et al., 2018) y SPARQL para mejorar la accesibilidad y el procesamiento de los datos LOD. Como se puede observar en la Figura 6, se produjo un notable aumento en *linked open data*.

Figura 5

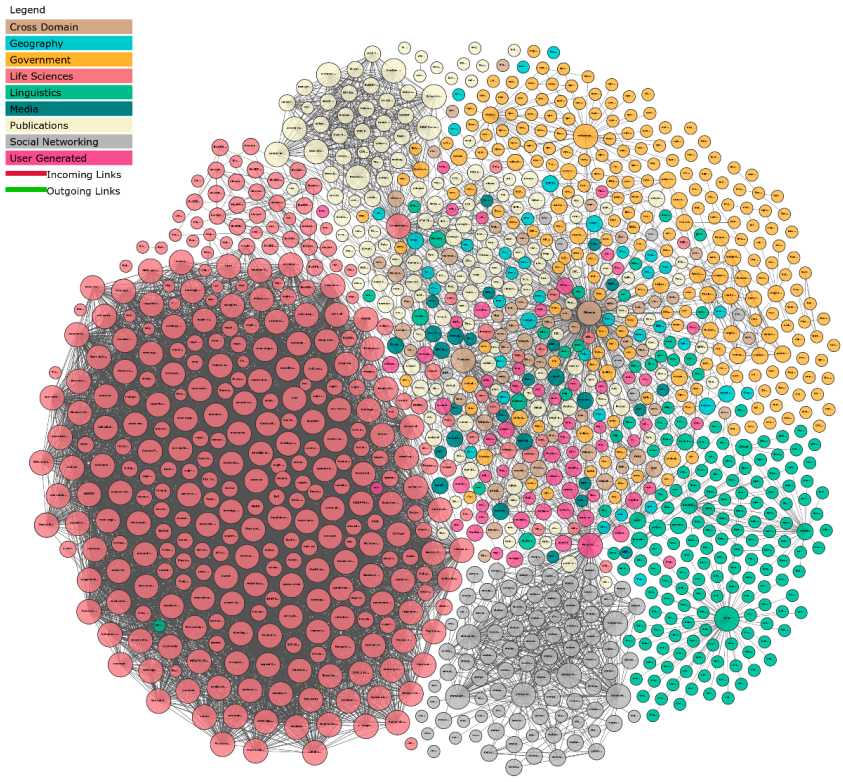
Crecimiento LOD Cloud 2014



Fuente: TLODC (2024a).

Figura 6

Crecimiento LOD Cloud 2017



Fuente: TLODC (2024a).

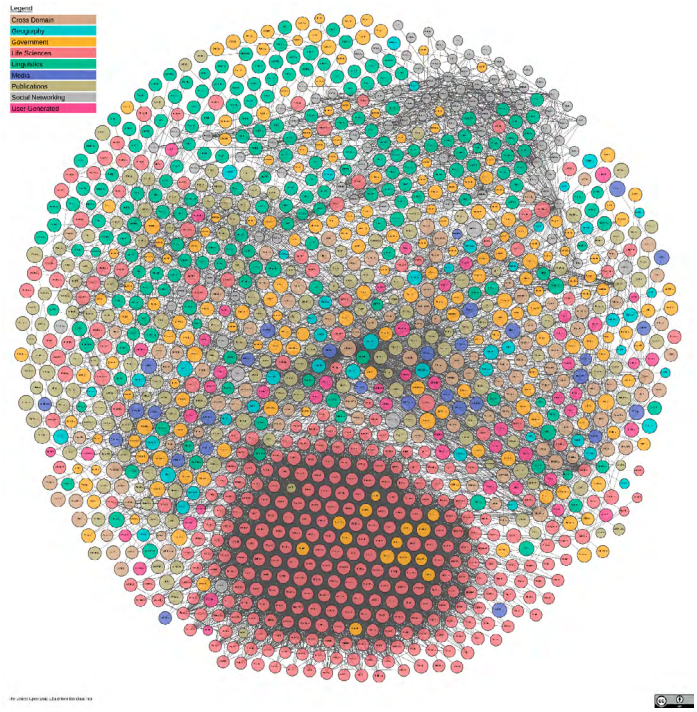
Para el 2018 se generan varias liberaciones de conjunto de datos pasando a 1231 *datasets*, siguiendo con los dominios definidos por LOD Cloud (véase Figura 7). Sin embargo, el dominio de ciencias sociales se identifica como el dominio que más contenido ha liberado, tendencia de crecimiento para este dominio evidenciada desde los gráficos de años previos. La interoperabilidad entre conjuntos de datos era un desafío importante, dado que el uso de LOD Cloud se concentraba principalmente en la investigación académica (Beek et al., 2020).

Durante el 2019 a la fecha, se lanzaron nuevas herramientas para facilitar la publicación y el consumo de datos LOD (Asprino et al., 2019). La comunidad de usuarios de LOD Cloud comenzó a crecer, con un mayor interés por parte de la industria. Pero a raíz de la pandemia de COVID-19, se buscó una aceleración por la adopción de tecnologías digitales; LOD Cloud no se vio beneficiada en sí. Sin embargo, la madurez de *linked open data* se evidencia: se desarrollaron nuevos casos de uso para LOD Cloud en áreas como la salud, la educación y el gobierno y se expandió a nuevos sectores como la agricultura, la energía y el turismo (Monaco et al., 2022).

Durante el 2022 se desarrollaron proyectos para integrar LOD Cloud con otras plataformas de datos como Wikidata (Wikidata, 2024) y Google Dataset Search. Es así que la comunidad de LOD Cloud se globalizó con la participación de actores de todo el mundo. Para el 2023 se avanza en la creación de un ecosistema LOD Cloud más robusto y sostenible. Se desarrollan nuevas herramientas y aplicaciones para facilitar el uso de LOD Cloud por parte de una audiencia más amplia (Maillot et al., 2023). LOD Cloud se consolida como una herramienta fundamental para la gestión y el análisis de datos en la era digital, donde se tiene a la fecha un total de 1314 *datasets* con 16308 *links* (véase Figura 8).

Figura 7

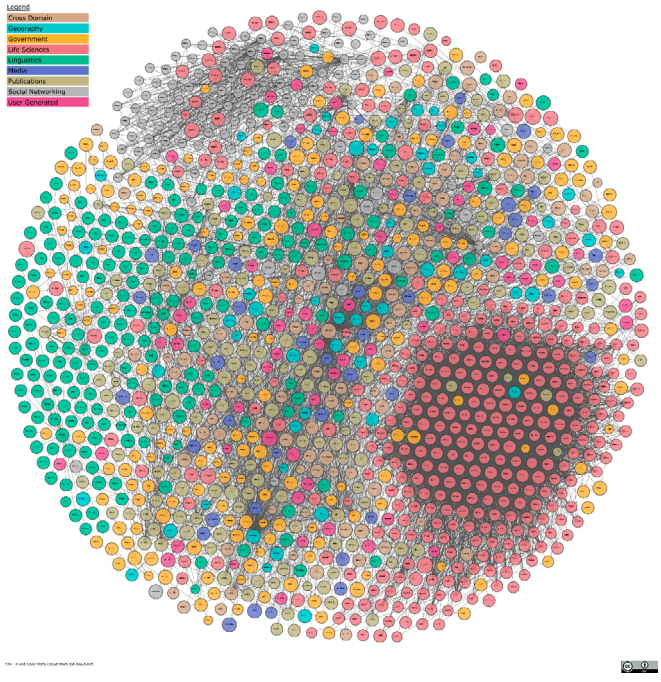
Crecimiento de LOD Cloud al 2018



Fuente: TLODC (2024a).

Figura 8

Crecimiento LOD Cloud 2022



Fuente: TLODC (2024a).

Desde este contexto, a continuación se hace un análisis breve del nivel de exposición de estos recursos.

Análisis de crecimiento LOD

En este apartado se plantea el método estadístico descriptivo, donde se eligen tres variables que se condensarán:

- Cantidad de *datasets*
- Cantidad de tripletas RDF
- Cantidad de *links* RDF

Se ha condensado en la Tabla 2 la información redactada en la sección anterior, que permite visualizar las cantidades en un periodo de tres años y un rango de tiempo de 15 años, 2008 a 2023.

La evolución del número de *datasets* y *links* en el proyecto LOD Cloud se presenta a continuación con una serie de gráficas, que pretenden facilitar el análisis del crecimiento del volumen de datos. En el caso de las tripletas, las fuentes consultadas difieren en su método de obtención de datos; mientras que Bizer & Berners-Lee (2008) y Bizer et al. (2011) hacen *crawling* con diferentes herramientas, los TLODC (2024, 2024a) se basan en la información recolectada por DataHub y hacen diferenciación entre *links* incorrectos. Por tanto, no tiene sentido hacer una gráfica para visualizar el crecimiento, ya que se vería bastante impreciso y carecería un poco de sentido analítico.

Primero analizaremos los *datasets*. La primera figura muestra que el número de *datasets* aumenta en todos los años en un buen número; sin embargo, la segunda gráfica muestra que entre el 2008 y el 2009 se presentó el menor incremento porcentual, que del 2011 al 2014 la tasa de crecimiento rebasó a las anteriores, que a finales del 2017 se duplicó y en los años siguientes se evidencia una desaceleración (véase Figura 9).

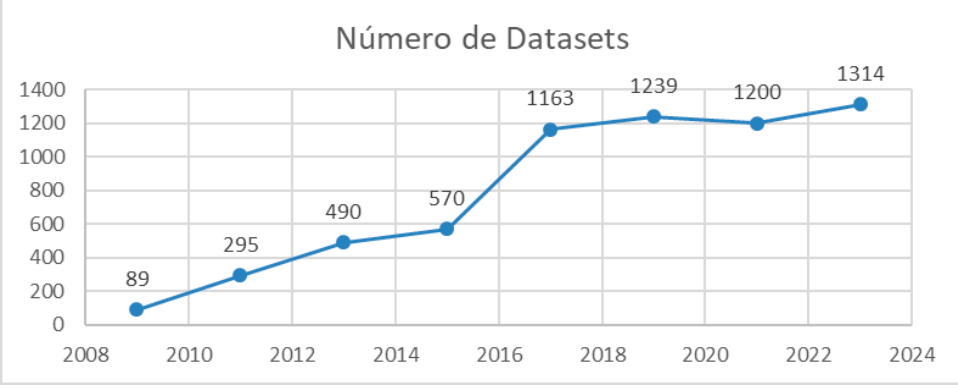
Tabla 2

Relación histórica de tripletas establecidas bajo formato RDF

Cantidad/año	2008	2011	2014	2017	2020	2023
Datasets	93	295	592	1163	1255	1314
Tripletas RDF	4,700,000	31,634,213	101,232,000	1,486 M	6,844 M	28,000 M
Links RDF	14,000,000	503,998,830	-	-	-	-

Figura 9

Histórico número de datasets 2008-2024



Por otra parte, los *links* RDF muestran un aumento mucho más pronunciado cada año. Teniendo en cuenta que el número de *links* RDF es la cantidad de enlaces entre *datasets*. Es lógico que cuando aumenta significativamente el número de *datasets* en el diagrama, también aumente, aunque no siempre de manera tan drástica, lo mismo con el número de *links* RDF, pues serán necesarios para conectar los *datasets* (véase Figura 10).

Respecto a la relación del crecimiento de los *datasets* y los *links* RDF, se presenta la Tabla 3, donde se relaciona el número de cada elemento presente desde el 2009 al 2023.

Otras investigaciones asociadas a LOD y fuentes de datos inteligentes (OSINT)

Existe un enorme potencial de recursos que se exponen como fuentes de datos abiertos en diversos escenarios, pasando por la reutilización de recursos educativos (Herrera-Cubides et al., 2020), la ciberseguridad (Pastor-Galindo et al., 2020), entre otros. Sin embargo, si no se cuenta con un adecuado tratamiento y proceso de transformación, no será posible extraer aquellos aspectos que permitan su uso para analizar y posteriormente utilizar como fuentes de datos inteligentes.

En la literatura podemos hallar diversos trabajos relacionados con OSINT y sus diferentes aplicaciones; en algunos de los más relevantes se pueden encontrar herramientas analizadas por CIA (2018), como Oryon, un navegador web diseñado para ayudar a los investigadores a realizar investigaciones de Open Source Intelligence. Oryon viene con docenas de herramientas preinstaladas y un conjunto selecto de enlaces catalogados por categoría. Otra herramienta de Facebook Graph Search (Sowsearch, 2024), permite extraer datos de cada persona y página en toda la red social. En el internet se encuentran múltiples fuentes de tipo *open-source* para diferentes tipos de búsqueda: videos, imágenes, textos, entre otros.

Desde hace varios años se han propuesto iniciativas para promover herramientas que puedan capturar recursos desde diversas áreas de conocimiento, como OSINT Framework (Nordine, 2024) que permite explorar, a través de una estructura de navegación, 32 tipos de categorías que se pueden consultar en diferentes fuentes de datos abiertos (véase Figura 11).

Figura 10

Relación número de links RDF registrados

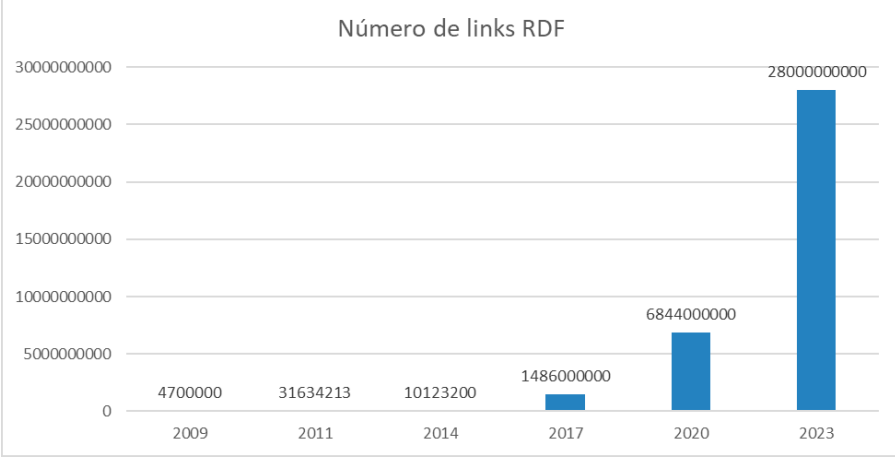


Tabla 3

Relación número de datasets *2009-2023*

	2009	2011	2014	2017	2020	2023
No. Datasets	93	295	592	1163	1255	1314
No. Links RDF	14,000,000	503,998,830	--	--	--	--

Figura 11

OSINT Framework

Fuente: Nordine (2024).

En el internet se encuentran herramientas de código abierto que se conectan a varios sitios web y verifican, por ejemplo, la presencia de los nombres de usuario en todos los sitios web a la vez. Algunas de estas herramientas son (Passi, 2018):

- Maltego: utilizada para investigación forense y de seguridad.
- Shodan: motor de búsqueda para hackers.
- Google Dorks: mejora la busque e indexación de resultados.
- The Harvester: información relacionada con correos y dominios.
- Metagoofil: metadata de documentos públicos.
- Check Usernames: verificar los nombres de usuario en una red social cualquiera.
- TinEye: buscar imágenes en la web.
- Searchcode: buscar código en la web.

Otras herramientas, aplicaciones y proyectos se pueden encontrar en Bielska et al. (2018), Hocke (2020) y AOS (2020). Algunas aplicaciones de herramientas OSINT se pueden ver en campos como (Pastorino, 2019):

- Búsquedas por ubicación: los comentarios subidos a la web son georreferenciados, por lo que herramientas como Geo Twitter u otras del proyecto Geo Social Footprint, son útiles para realizar búsquedas de noticias o posteos, dada una ubicación.
- Palabras clave: cuando se definen palabras clave, es de vital importancia tener en cuenta el idioma y la jerga utilizada. Sitios como Newspaper Map proveen diarios locales de diferentes regiones de todo el mundo, con lo que se pueden construir las palabras clave requeridas.
- Datos laborales, documentos de identidad, entre otros: individuos sin interacción en redes sociales, cuentas bancarias o tarjetas de crédito, pueden ser encontrados a través de páginas gubernamentales por medio de su documento de identidad, inscripción tributaria, servicios públicos o incluso infracciones de tránsito.
- Generación de identidades para la investigación: aplicaciones como Fake Name Generator permite crear datos de un individuo y Thispersondoesnotexist.com, crear fotos falsas a partir de la inteligencia artificial. Aplicaciones como PostCron permite configurar posteos automatizados, simulando actividad de estos perfiles, haciéndolos parecer reales.

En general, los procesos de captura de datos que llevan a cabo estas herramientas, permiten tener una variedad de formatos y usos de la información proporcionada por las fuentes. Por tanto, su uso en la toma de decisiones puede variar en campos de índole gubernamental, organizaciones internacionales, agencias militares y judiciales, corporaciones de negocios, organizaciones que mitigan aspectos de ciberseguridad, así como las mismas organizaciones criminales y terroristas, entre otros. Por consiguiente, resultan ser fuentes de información valiosa, que permiten ser utilizadas en áreas de seguridad para potenciar los resultados. Pastor-Galindo et al. (2019) describen el estado actual de OSINT y hacen una revisión exhaustiva del paradigma, centrándose en los servicios y técnicas que mejoran el campo de la seguridad cibernética.

Conclusiones

El análisis llevado a cabo permite concluir que la pandemia surgida en el 2020 a causa del *SARS-CoV-2* ha impactado directamente en el proceso de vinculación de datos abiertos por LOD Cloud y las distintas iniciativas asociadas a LOD*, lo cual afecta en su aplicación en las fuentes de datos inteligentes (OSINT), identificando un estancamiento en el proceso de liberación de datos (datos abiertos) y en el nivel de madurez de su vinculación.*

Iniciativas como LOD Cloud son de gran importancia en la actualidad, ya que permiten identificar el crecimiento de esta iniciativa para facilitar una categorización de toda la información en internet. La vinculación de datos es un importante avance hacia el ideal de la web semántica. Sin embargo, han surgido distintos retos asociados con la gran cantidad de número de *links* y *datasets*; las tecnologías utilizadas para establecer esta vinculación serán fundamentales para facilitar la navegación en diversas áreas de conocimiento, facilitando a los motores de búsqueda obtener resultados semánticamente correctos.

Desde sus comienzos, LOD Cloud se ha convertido en una plataforma esencial para la web semántica, utilizada en una amplia variedad de aplicaciones y áreas de conocimiento, desde la investigación científica hasta la toma de decisiones en los sectores público y privado. La comunidad de LOD Cloud continúa creciendo y colaborando para hacer de la web semántica una realidad. Algunos de los puntos más relevantes identificados incluyen una madurez en diferentes aspectos relacionados con:

- El crecimiento exponencial en la cantidad de conjuntos de datos disponibles.
- El lanzamiento de nuevas herramientas y aplicaciones para facilitar el uso de LOD Cloud.
- La consolidación de la comunidad de LOD Cloud a nivel global.
- La adopción de LOD Cloud por diversos sectores de la industria.

Como trabajos futuros, se considera importante analizar las fuentes de datos actuales que se tienen como datos vinculados según especificaciones LOD, con el objetivo de explotar los beneficios enmarcados dentro de las fuentes de datos abiertos (OSINT). Esto permitirá mejorar la riqueza y semántica en la que se exponen estos datos, de tal manera que se puedan estructurar y consumir para facilitar procesos de análisis y extracción de información inteligente.

Durante el análisis de crecimiento realizado, se ha evidenciado la aceptación que ha recibido la iniciativa LOD Cloud hasta el 2022, lo que ha motivado a cientos de personas a darse a la tarea de vincular contenidos libres existentes en la red. No obstante, de acuerdo con el análisis de este crecimiento desde el 2022 a la fecha, parece existir un leve estancamiento en cuanto a la vinculación de más fuentes de datos, lo que justifica abrir un debate que permita identificar las limitaciones y retos a los que está expuesta esta iniciativa, y, por tanto, poder determinar si el nivel de madurez que presenta actualmente, limitaría su aplicación dentro de las fuentes de datos abiertos inteligentes (OSINT).

Referencias

AOS (Awesome Open Source). (2020). *The Top 146 Osint Open Source Projects*. Awesome Open Source. https://awesomeopensource.com/projects/osint

Asprino, L., Beek, W., Ciancarini, P., Harmelen, F. V., & Presutti, V. (2019). Observing LOD using equivalent set graphs: It is mostly flat and sparsely linked. En *International Semantic Web Conference* (pp. 57-74). Springer. https://doi.org/10.1007/978-3-030-30793-6 4

Beek, W., Raad, J., Acar, E., & van Harmelen, F. (2020). MetaLink: A travel guide to the LOD cloud. En *The Semantic Web: 17th International Conference*, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings 17 (pp. 481-496). Springer International Publishing.

Bielska, A., Anderson, N., Benetis, V., & Viehman, C. (2018). *Open source intelligence tools and resources handbook*. I-Intelligence. https://www.i-intelligence.eu/wp-content/uploads/2018/06/OSINT_Handbook_June-2018_Final.pdf

Bikakis, N., Tsinaraki, Ch., Gioldasis, N., Stavrakantonakis, I., & Christodoulakis, S. (2013). The XML and semantic web worlds: Technologies, interoperability and integration. A survey of the state of the art. En I. Anagnostopoulos, M. Bieliková, P. Mylonas, & N. Tsapatsoulis (Eds.), *Semantic hyper/multimedia adaptation. Studies in computational intelligence* (vol. 418). Springer. https://doi.org/10.1007/978-3-642-28977-4_12.

Bizer, C., & Berners-Lee, T. (2008). ChristianLinked Data: Principles and state of the art. *17th International World Wide Web Conference W3C Track @ WWW2008*. Beijing, China. http://www.w3.org/2008/Talks/WWW2008-W3CTrack-LOD.pdf

Bizer, C., Heath, T., & Berners-Lee, T. (2011). Linked data: The story so far. En *Semantic services, interoperability and web applications: Emerging concepts* (pp. 205-227). IGI Global.

Carcaño, F. (2018). *What is OSINT and what are open sources? FCD intelligence*. https://www.fcd-intelligence.com/2018/09/que-es-osint-y-que-son-fuentes-abiertas/

Herrera-Cubides, J. F., Gaona García, P. A., & Sánchez Alonso, S. (2020). Open-source intelligence educational resources: A visual perspective analysis. *Applied Sciences*, *10*(21), 7617.

Hocke, R. (2020). *Internet tools and resources for open-source intelligence (OSINT).* http://www.onstrat.com/osint/

LOD2. (2024). *Information and communication technologies*. UK Public Data. LOD2 Stack. https://lod2.eu/

LOD Cloud Draw. (2024). *LOD Cloud Draw*. https://github.com/lod-cloud/lod-cloud-draw

Maillot, P., Corby, O., Faron, C., Gandon, F., & Michel, F. (2023). IndeGx: A model and a framework for indexing RDF knowledge graphs with SPARQL-based test suits. *Journal of Web Semantics*, *76*, 100775. https://www.sciencedirect.com/science/article/abs/pii/S1570826823000045

Monaco, D., Pellegrino, M. A., Scarano, V., & Vicidomini, L. (2022). Linked open data in authoring virtual exhibitions. *Journal of Cultural Heritage*, *53*(22), 127-142. https://doi.org/10.1016/j.culher.2021.11.002

Nordine J. (2024). The osint framework. https://osintframework.com. (Ultimo acceso: 1 Julio 2024)

Passi, H. (2018). *Top 10 popular open source intelligence (OSINT) tools*. GreyCampus. https://www.greycampus.com/blog/information-security/top-open-source-intelligence-tools

Pastor-Galindo, J., Nespoli, P., Gómez Marmo, F., Martínez Pérez, G. (2019). OSINT is the next internet goldmine: Spain as an unexplored territory. *Conference: V Jornadas Nacionales de Investigación en Ciberseguridad*. https://www.researchgate.net/publication/333703698_OSINT_is_the_next_Internet_goldmine_Spain_as_an_unexplored_territory

Pastor-Galindo, J., Nespoli, P., Mármol, F. G., & Pérez, G. M. (2020). The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends. IEEE access, 8, 10282-10304.

Pastorino, C. (2019). *Técnicas y herramientas OSINT para la investigación en internet*. Welivesecurity by ESET. https://www.welivesecurity.com/la-es/2019/10/07/tecnicas-herramientas-osint-investigacion-internet/

Peis, E., Herrera Viedma, E., & Morales del Castillo, J. (2009). Aproximación a la web semántica desde la perspectiva de la documentación. *Investigación Bibliotecológica. Archivonomía, Bibliotecología e Información*, *21*(43). https://doi.org/10.22201/iibi.0187358xp.2007.43.4138

Pune, M. (2020). *Open source intelligence (OSINT). Market research report-global forecast to 2023—Market analysis, scope, stake, progress, trends and forecast to 2023. Market research future*. https://www.marketresearchfuture.com/reports/open-source-intelligence-market-4545

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). *Estado del LOD Cloud 2014*. Universidad de Mannheim. http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/

Sowsearch (2024). Facebook Search Filter tool. Website: https://www.sowsearch.info/, (Último Acceso: Julio. 1, 2024).

TLODC. (2024). *The Linked Open Data Cloud*. https://lod-cloud.net/

TLODC. (2024a). *Estado de LOD Cloud en 2009*. http://lod-cloud.net/versions/2009-03-05/lod-cloud.png

Wikidata. (2024). https://www.wikidata.org/?uselang=es

Wylot, M., Hauswirth, M., Cudré-Mauroux, P., & Sakr, S. (2018). RDF data storage and query processing schemes: A survey. *ACM Computing Surveys (CSUR)*, *51*(4), 1-36.