C.D.M Cuadernos de Documentación Multimedia

Vol. 25 Año 2014

ISNN: 1575-9733

Linked Open Media Data: La tecnología de datos vinculados al servicio de los medios de comunicación

Ana B. Ríos Hilario

Recibido: 14 de noviembre de 2014

Aceptado: 2 de diciembre de 2014

Resumen

Se analiza la aplicación de las tecnologías de *linked open data* (LOD) en el entorno de los medios de comunicación a partir del estudio de los conjuntos de datos, específicos de este ámbito, presentes actualmente en la "nube" de los datos vinculados. Mediante la aplicación de una metodología fundamentalmente de carácter descriptivo, el artículo se estructura en torno a los siguientes tres grandes apartados: en el primero se analiza el concepto de LOD; a continuación se realiza el estudio propiamente dicho de los *dataset* referentes a los medios de comunicación, y finalmente, se examinan cada uno de estos conjuntos presentes en el catálogo *Mannheim.* Concluye el artículo destacando las principales características obtenidas y sugiriendo posibles líneas de investigación sobre esta misma temática.

Palabras clave

Linked open data (LOD), Diagrama de la nube LOD, Conjuntos de datos, Medios de comunicación, *Mannheim Linked data Catalog*

Linked Open Media Data: Linked data technology at the service of the media

Abstract

The application of technologies linked open data (LOD) is analyzed in the environment of the media through the study of datasets, specific to this area, currently present in LOD cloud diagram. Through the application of a methodology mainly descriptive the article is structured around three main sections: in the first the concept of LOD is analyzed; then the dataset study concerning the media is performed, and finally, each of these sets present in the Mannheim catalog are discussed. It is concluded highlighting the main features found and suggesting possible researches on this topic.

Keywords

Linked open data (LOD), LOD cloud diagram, Dataset, Media, *Mannheim Linked data Catalog*

http://dx.doi.org/10.5209/rev_CDMU.2014.v25.47470

INTRODUCCIÓN

La tecnología actual y en parte también las nuevas formas de comunicación como es el caso de las redes sociales ha propiciado que ya no estemos interesados en un documento en su totalidad, como un libro o un periódico, ni siquiera en sus partes componentes, lo que sería un capítulo o artículo determinado, sino que de esos recursos el lector o investigador puede requerir una determinada tabla o un gráfico concreto. Es así como hemos pasado de solicitar documentos a demandar datos.

Por tal motivo las organizaciones de todo tipo han decidido publicar sus datos en abierto, lo que se conoce con el nombre de *open data*.

Por su parte, la web semántica permite pasar de una web de documentos a una web de datos en la cual unos datos se conectan o enlazan con otros. Se abre así la posibilidad de enlazar conjuntos de datos (*datasets*) con otros conjuntos y en última instancia datos con datos de acuerdo con una serie de principios y modelos de interrogación bien establecidos (*Linked data web*, 2014). Estamos por lo tanto ante un nuevo fenómeno, el que hace referencia a los datos abiertos y vinculados, más conocido por su acrónimo inglés LOD (*linked open data).*

Son numerosos los casos y ejemplos del empleo de la tecnología de los datos vinculados en cualquier ámbito pero, centrándonos en el caso particular de los medios de comunicación, diremos que Tim Berners Lee en una de sus múltiples conferencias expuso como el diario británico *The Times* utilizó la información pública del Gobierno británico para generar un tipo de información que antes no existía: un mapa de accidentes de bicicletas. Los datos así recolectados fueron dispuestos en un mapa de Google lo que permitía a los lectores conocer las zonas donde hay más accidentes de bicicletas o saber dónde ocurrió cada uno de forma individual. Lo importante de este ejemplo es que el periódico tomó los datos de la web del Gobierno británico (gov.uk) (Mazzo, 2010).

A partir de lo anteriormente expuesto, consideramos que sería interesante realizar un estudio que analizara la disposición actual y representación de los conjuntos de datos pertenecientes a los diferentes medios de comunicaciones existentes en la "nube de los datos vinculados". De este propósito general podemos a su vez definir los siguientes objetivos específicos.

- Definir brevemente los distintos conceptos asociados a la tecnología de datos enlazados presentes en su propia de denominación: *linked*, *open*, *data* (LOD);
- Explicar el diagrama de la nube de datos vinculados, centrándonos en su definición, objetivos y formación.
- Analizar de modo detallado las características generales de los *dataset* pertenecientes a la categoría específica de los *media*, haciendo referencia a las siguientes variables: conjuntos de datos disponibles y cumplimientos de una serie de buenas prácticas en lo referente a la interconexión, utilización de vocabularios y asignación de metadatos.

Vol. 25 Año 2014
ISNN: 1575-9733

- Descripción de los conjuntos de datos dispuestos en el *Mannheim Linked data Catalog* y etiquetados como medios de comunicación. Tales conjuntos los hemos subdividido a su vez en los siguientes grupos: películas, música, programas de televisión y radio y medios de comunicación impresos.

Para cumplir con estos objetivos hemos aplicado una doble metodología. En primer lugar, para contextualizar nuestra exposición hemos recurrido al estudio de diversas fuentes documentales que nos han permitido definir los principales conceptos presentes en la tecnología de *linked data* (LD). En segundo lugar, para el examen específico de los conjuntos de datos hemos aplicado una metodología de carácter descriptivo a través del análisis de dos recursos fundamentales: el propio diagrama de la nube de datos vinculados y el catálogo *Mannheim* anteriormente mencionado. Para el estudio y comprensión de la "nube" nos ha sido de mucha utilidad la consulta del documento titulado *Adoption of the linked data best practices in diferent topical domains* (Schmachtenberg; Bizer; Paulheim, 2014) y su versión abreviada disponible en el recurso *State of the LOD cloud 2014.*

En consonancia con los objetivos propuestos y la metodología empleada el artículo se estructura en tres grandes apartados. El primero, referente a la definición del término de LOD; el segundo centrado específicamente en el estudio de los *dataset* propios de los medios de comunicación; y el tercer punto, relativo al examen de cada uno de estos conjuntos presente en el catálogo *Mannheim.* Esta última parte nos parece realmente interesante al mostrarnos ejemplos concretos de medios de comunicación que aplican la tecnología de LD. Se ha intentado realizar una breve descripción de cada uno de ellos destacando los más importantes e interesantes desde esta perspectiva. Termina el artículo con una serie de conclusiones en las que se subrayan las características más importantes halladas en el análisis y en las que se señalan una serie de futuras líneas de investigación que sería interesante desarrollar.

LINKED OPEN DATA (LOD)

Los conceptos data, open y linked

En los últimos tiempos, la palabra "data" aparece con frecuencia en artículos científicos y de divulgación, generalmente, acompañado de otro término que delimita su significado como puede ser *open data*, del que posteriormente hablaremos, *data mining* y el más reciente *big data*. Además, bajo este concepto se agrupa toda una serie de datos de muy diverso tipo. Es así como, Hernández y García (2013) señalan que gracias a las tecnologías de la información podemos recopilar datos de carácter personal; datos sobre el contexto que vivimos, y datos sobre objetos y productos. Cada vez es más amplio el número de personas y organizaciones que están contribuyendo a este "diluvio de datos" al optar por compartir su información con los demás (Heath y Bizer, 2011). Sirvan de ejemplo instituciones tan importantes como Amazon, los

Vol. 25 Año 2014
ISNN: 1575-9733

organismos públicos como el Gobierno de los Estados Unidos, instituciones bibliotecarias como *Europeana* y la *Library of Congress* e iniciativas de investigación en diversas disciplinas científicas. Y, por supuesto, en los medios de comunicación grupos como la *BBC* y periódicos tan relevantes como el *New York Times* de los que hablaremos más adelante. En este contexto aparece el concepto de *data journalism* que hace referencia a una especialidad del periodismo que refleja el creciente papel que los datos numéricos tienen en la producción y distribución de la información en la era digital (*Data journalism handbook*, 2012).

Por otro lado, el adjetivo "open" asociado a otros términos, ha dado lugar a los denominados "movimientos abiertos". El concepto de *open* se incluye dentro de una expresión más amplia que es la de conocimiento abierto. Según el *Open Definition Advisory Council* esta expresión haría referencia tanto al "contenido incluido en música, películas y libros; los datos de carácter científico, histórico, geográfico; o cualquier otro tipo información gubernamental y de otras administraciones públicas". Este mismo organismo expresa las condiciones de distribución que debe cumplir una obra para que sea considerada abierta. En concreto enumera las siguientes 11 pautas: acceso, redistribución, reutilización, ausencia de restricciones tecnológicas, reconocimiento, integridad, sin discriminación de personas o grupos, sin discriminación de ámbitos de trabajo, distribución de la licencia, la licencia no debe ser específica de un paquete y la licencia no debe restringir la distribución de otras obras.

La lista de expresiones que abarca el término *open* es cada vez más amplia aunque tenemos que decir que la primera de ellas fue la de *open source* que surge a finales de la década de los 90. El código abierto no sólo significa el acceso al código fuente (Open Source Iniciative, 1998) sino que deben cumplir toda una serie de criterios especificados en la propia iniciativa.

Dentro de este contexto destaca también el o*pen access* que se define como la disponibilidad gratuita de la literatura en internet que permite que cualquier usuario pueda leer, descargar, copiar, imprimir, distribuir, buscar y enlazar información sin barreras financieras, legales o técnicas (Budapest open access initiative, 2002). En este sentido podemos decir que *open access* propicia el libre acceso a la información y conocimiento a través de internet sin barreras económicas y sin restricciones derivadas de los derechos de copyright (Ferreras, 2011). En este contexto, son los propios autores quienes definen los derechos que otorgan a sus trabajos que generalmente se realizan a través de licencias *creative commons* (Martín y Angelozzi, 2013). En este mismo entorno se hallaría también el término *open data* que pasaremos a definir a continuación.

Finalmente, el último de los términos en este ámbito no se define por sí mismo sino que va unido en primer lugar a la expresión *data*, y en segundo lugar, a la suma de *open* y de *data*.

Vol. 25 Año 2014
ISNN: 1575-9733

Las iniciativas open data, linked data y linked open data

La iniciativa *open data* está estrechamente relacionada y vinculada a la concepción de gobierno abierto, siendo su "filosofía la del acceso abierto a determinados datos sin restricciones de copyright" (Ferrer; Peset; Aleixandret, 2011, p. 162). En este sentido podemos definir *open data* como "un movimiento que promueve la liberación de datos, generalmente no textuales y en formatos reutilizables como CSV (comma separated values), procedentes de organizaciones" (Peset; Ferrer; Subirats, 2011).

Conviene aclarar que *open data* no es la mera disponibilidad de los datos en la red, es decir, su publicación en internet de modo que los datos puedan leerse y descargarse. En este sentido Hernández y García (2013, p. 260) afirman que "para que sean realmente abiertos sí deben estar disponibles en internet, preferiblemente para ser descargados, pero también deben poseer algún tipo de licencia legal para poderlos utilizar, reutilizar y redistribuir, mezclándolos incluso con otros datos, sujetos como mínimo a la "atribución" (reconocimiento de la autoría, quién lo ha hecho), o al "compartir igual" (que la explotación que se haga de esos datos –incluyendo las obras derivadas– mantengan la misma licencia al ser divulgadas)".

En cuanto a las pautas a seguir para la publicación de estos datos, fueron establecidas por el World Wide Web Consortium (W3C) en el documento titulado *Publishing open government data* (2009). Podemos resumir estas recomendaciones en los siguientes puntos:

1. Publicación de los datos en bruto y en un formato que permita el uso automatizado (xml, rdf y csv).
2. Creación de un catálogo en línea de dichos datos para que el usuario pueda conocer que datos se han publicado.
3. Disposición de los datos de modo legible tanto por las personas como por las máquinas.

Berners Lee, desarrolló en 2010 un sistema de clasificación de clasificación de 5 estrellas "con el fin de animar a la gente –especialmente a los propietarios de los datos del gobierno– en el camino hacia los buenos datos vinculados"

Vol. 25 Año 2014
ISNN: 1575-9733

★	Datos disponibles en la web, en cualquier formato, pero con una licencia abierta para ser *open data* (por ejemplo, un pdf)
★★	Datos en formato estructurado (por ejemplo un Excel)
★★★	Datos en un formato no propietario: por ejemplo, ficheros de datos separados por comas (*comma separated values*, CSV)
★★★★	Uso de un estándar W3C (RDF y SPARQL) para identificar las cosas, de modo que otras personas puedan apuntar a esas cosas.
★★★★★	Enlazar unos datos con otras fuentes de datos para dotarlos de contexto

Figura 1. Elaboración propia a partir de la información de Tim Bernes Lee, 2010

Si queremos conocer de modo preciso el significado del término *linked data* debemos acudir a la página oficial, denominada del mismo modo, en la que se define este concepto como "la utilización de la Web para conectar los datos relacionados que no estaban vinculados previamente, o el uso de la Web para disminuir los obstáculos en la conexión de los datos actualmente vinculados mediante otros métodos". Dicho de otro modo, LD sería "la publicación de datos estructurados que permiten la conexión y enriquecimiento de los metadatos" (Ríos, 2014), de tal forma que "diferentes representaciones de un mismo contenido puedan encontrarse y enlazarse entre recursos relacionados" (*Europeana Linked Open Data*, 2014).

El término datos vinculados se refiere, por lo tanto, a un conjunto de buenas prácticas para la publicación y la interconexión de datos estructurados en la Web. Para hablar de LD los datos deben publicarse de acuerdo con los principios diseñados para facilitar los vínculos entre los conjuntos de datos, elementos y vocabularios controlados (Berners Lee, 2006). Estas prácticas fueron presentadas en 2006 por Tim Berners Lee y se han dado a conocer como los principios de *linked data*. Tales principios son los siguientes:

1. Usar URIs (*uniform resource identifiers*) para identificar los recursos de forma unívoca.
2. Usar URIs http para que la gente pueda acceder a la información del recurso.
3. Ofrecer información sobre los recursos usando RDF.
4. Incluir enlaces a otros URIs, facilitando el vínculo entre los distintos datos distribuidos en la web.

Es decir, este conjunto de buenas prácticas o recomendaciones que se incluyen bajo este concepto van a "permitir, exponer, compartir y conectar conjuntos de datos, información y conocimiento en

Vol. 25 Año 2014
ISNN: 1575-9733

la web semántica mediante la utilización de URIs para la identificación de los recursos y de estándares como RDF para la descripción de las mismas" (Ríos; Ferreras; Martín, 2014)

Según Mitchell (2013, p. 12) "LOD comprende dos conceptos distintos, el primero es que los datos publicados en la Web deberían conectarse fácilmente con información ("enlazados") y al hacer esto debería ser accesible tanto para las computadoras como para los humanos ("datos"). El segundo concepto clave para LOD es que para que los datos se vinculen y se reutilicen, deben estar abiertos y libres de las restricciones de derechos y copyright ("abiertos").

Aunque en algunos contextos se suele asimilar ya *linked data* con *linked open data*, no todos los datos enlazados son datos abiertos y no tienen por qué serlo. Sólo lo son si se liberan bajo acuerdos de licencia que permiten su libre reutilización. Es decir, mientras *linked data* alude a la interoperabilidad técnica de los datos, *open data* hace referencia a su interoperabilidad legal.

LA NUBE DE DATOS Y SU RELACIÓN CON LOS MEDIOS DE COMUNICACIÓN

Un paseo por la nube

En la página web denominada *LOD cloud diagram* (diagrama de la nube LOD) se muestran los conjuntos de datos que se han publicado en el formato de datos vinculados y que son recogidos por los colaboradores del proyecto Linking Open Data y de otra serie de personas y organizaciones. El proyecto Linking Open Data (2014) tiene como objetivo primordial "extender la Web como un bien común de datos mediante la publicación de varios conjuntos de datos abiertos como RDF en la Web y mediante el establecimiento de enlaces RDF entre elementos de datos de diferentes fuentes". Dicho diagrama se basa en los metadatos recogidos y comisariados por los colaboradores del Data Hub, así como en los metadatos extraídos de un rastreo de la web *linked data* que se realizó en abril de 2014. Podemos definir a su vez Data Hub como "un registro de datos en el que se puede compartir información sobre paquetes de datos de cualquier tipo y describirlos de forma colaborativa" (Isaac; Waites; Young; Zeng, 2011).

Vol. 25 Año 2014
ISNN: 1575-9733

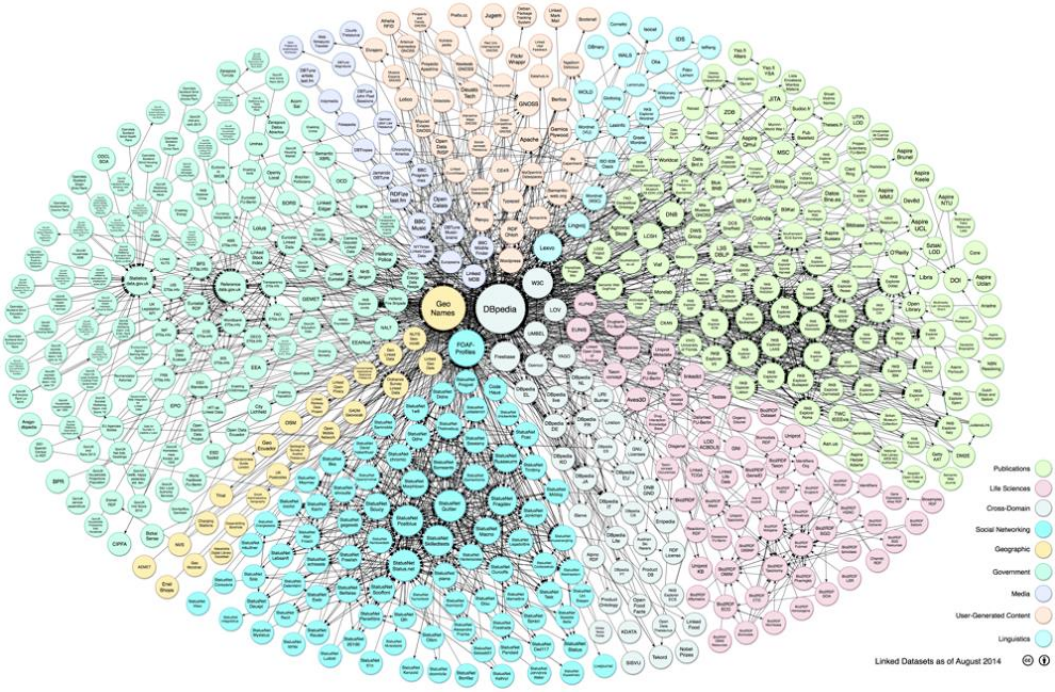

Figura 2. Diagrama de la nube LOD

Fuente: http://lod-cloud.net/

En la nube LOD podemos encontrarnos con dos páginas independientes con estadísticas acerca de estos conjuntos de datos dependiendo del momento en el que se captura la imagen.

- El estado de la nube en 2014: documento que presenta información sobre la estructura y el contenido del subconjunto rastreado de la nube LOD en abril de 2014.

- El estado de la nube 2011: documento que proporciona información sobre los conjuntos de datos vinculados que fueron catalogadas en el Data Hub en septiembre de 2011.

El diagrama generado en el año 2011 se basa exclusivamente en los metadatos del grupo Lodcloud presente en el Data Hub, mientras que la última versión (2014) se basa en los metadatos extraídos a partir de dos fuentes. En concreto, de los 570 conjuntos de datos que actualmente figuran en el diagrama:

- 374 fueron descritas por los propios proveedores de datos en el Data Hub.

- 196 conjuntos de datos fueron descubiertos por un rastreo de la web *linked data* realizada en abril de 2014.

Vol. 25 Año 2014
ISNN: 1575-9733

La interpretación del diagrama es sencilla. La imagen nos muestra los conjuntos de datos publicados en formato *linked data* y que se entrelazan con otros conjuntos de datos de la nube. El tamaño de los círculos se corresponde al número de tripletas en cada conjunto de datos. Generalmente, los números son proporcionados por los editores de los conjuntos de datos y en ocasiones son cálculos aproximados.

Los conjuntos de datos referentes a los medios de comunicación se identifican por el color azul claro y se sitúan en el margen superior izquierdo, casi en la parte central, del diagrama.

Las flechas muestran la existencia de al menos 50 enlaces entre dos conjuntos de datos. Un enlace, según el propósito de la nube, es una tripleta RDF donde sujeto y objeto URIs están en los espacios de nombres de diferentes conjuntos de datos.

La dirección de las flechas indica el conjunto de datos que contiene los enlaces, por ejemplo, una flecha de A a B significa que el conjunto A contiene tripletas RDF que utilizan los identificadores de B. Las flechas bidireccionales generalmente indican que los enlaces se reflejan en ambos conjuntos de datos. El espesor se corresponde con el número de enlaces.

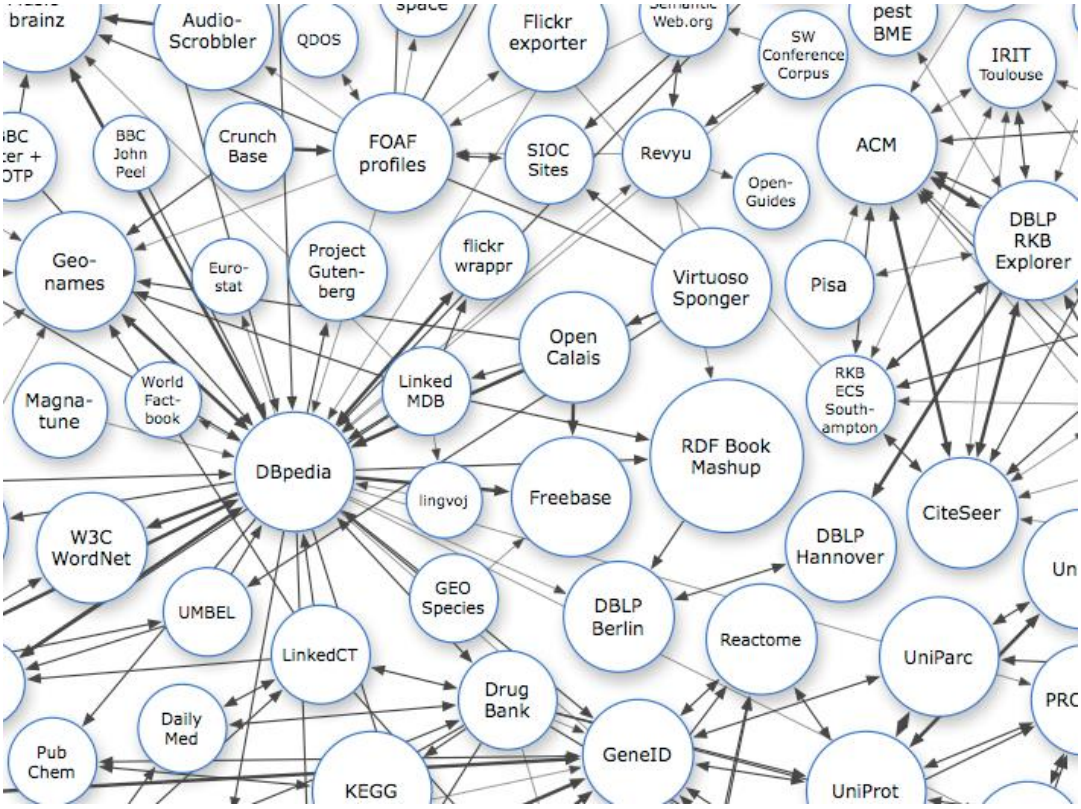

Figura 3. Parte del diagrama LOD

Fuente: http://linkeddata.org/

La nube de datos de los medios de comunicación

La tecnología LD se está empleando para compartir datos que abarcan una amplia gama de dominios de diferentes áreas. La siguiente tabla presenta una visión general de los 1.014 conjuntos de datos que se identificaron en el último rastreo clasificados por categorías temáticas.

Tema	Conjunto de datos	%
Gobierno	183	**18,05%**
Publicaciones	96	**9,47%**
Ciencias de la vida	83	**8,19%**
Contenido generado por el usuario	48	**4,73%**
Dominios cruzados	41	**4,04%**
Medios	22	**2,17%**
Geografía	21	**2,07%**
Web social	520	**51,28%**
TOTAL	**1014**	

Tabla I. Conjuntos de datos por categorías temáticas

Fuente: http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/

La categoría denominada medios de comunicación comprende los conjuntos de datos que proporcionan información sobre películas, música, programas de televisión y de radio, así como medios de comunicación impresos. Conjuntos de datos destacados dentro de esta categoría son los conjuntos de música dbtune.org, el conjunto de datos del periódico New York Times y los conjuntos de datos de los programas de radio y televisión de la BBC (Schmachtenberg; Bizer; Paulheim; 2014)

Como puede observarse en la tabla anterior de las 8 materias identificadas, la categoría "medios de comunicación" ocupa la penúltima posición, con 22 conjuntos de datos identificados, lo que se corresponde con el 2,17% del espacio en la nube. Destaca muy por encima del resto los conjuntos de datos relativos a la web social (520) seguido de lejos por los datos gubernamentales.

Si queremos realizar un análisis más exhaustivo de la realización del estado de la nube ejecutado en abril de 2014 tendremos que consultar el texto titulado *Adoption of the linked data best practices in diferent topical domains* (Schmachtenberg; Bizer; Paulheim, 2014) o el resumen del mismo que se establece en el documento *State of the LOD cloud 2014*.

De este documento es importante resaltar el apartado 4 destinado a una serie de buenas prácticas y que a continuación vamos a analizar detalladamente, ya que consideramos que no sólo es importante ver el estado actual de los conjuntos datos vinculados de los medios de comunicación y su posición con respecto a las otras áreas, sino que también esta radiografía podrá ser interesante y útil para aquellos medios que estén interesados en publicar sus datos como LOD.

LD parte de la base de que "los editores de datos soporten aplicaciones para el descubrimiento e integración de los datos a través del cumplimiento de un conjunto de buenas prácticas en las áreas de la vinculación, uso de vocabularios y provisión de metadatos" (Schmachtenberg; Bizer; Paulheim, 2014). En el citado documento se expone en qué medida los datos analizados implementan estas buenas prácticas.

Comenzando por el primer punto, el de la interconexión, mediante los enlaces RDF, los proveedores de datos conectan sus conjuntos de datos en un solo gráfico global que se puede navegar mediante las aplicaciones correspondientes y permitir el descubrimiento de datos adicionales siguiendo los enlaces RDF. En total, el 56,11% de los conjuntos de datos rastreados enlazan a al menos a otro conjunto de datos. De las diferentes tablas que se van presentando en este apartado, a nosotros nos interesa por su relación con nuestro objetivo, la última que hace referencia a los predicados más utilizados para la interconexión por categoría. En el caso de los medio los tres predicados más utilizados son los que se recogen en la siguiente tabla.

Categoría	Predicado	Uso
Medios	owl:sameAs	81,25%
	rdfs:seeAlso	18,75%
	foaf:based near	18,75%

Tabla II. Predicados más utilizados para la interconexión

El siguiente apartado hace referencia a los vocabularios empleados, es así como se especifica que "con el fin de hacer más fácil para las aplicaciones entender *linked data*, los proveedores de datos deben utilizar para representar los datos términos de vocabularios ampliamente desarrollados siempre que sea posible" (Schmachtenberg; Bizer; Paulheim, 2014). El documento establece una diferencia entre los vocabularios propietarios y los vocabularios de-referenciales.

Se define un vocabulario como no propietario si hay por lo menos dos conjuntos de datos que utilizan dicho vocabulario. De los 649 vocabularios encontrados, 378 (58,24%) son vocabularios propietario de acuerdo con esa definición, mientras que 271 (41.76%) son no propietario.

En cuanto a los tres vocabularios propietarios más utilizados en nuestra categoría son los que se recogen a continuación: *Friend of a Friend* (foaf), *Dublin Core Metadata Iniciative Terms* (dcterm), *Music Ontology* (mo).

Categoría	Vocabulario	Uso
Medios	foaf	75,67%
	dct	54, 05%
	mo	18,91%

Tabla III. Utilización de vocabularios propietarios

Para los vocabularios propietarios es esencial que sean de-referenciales y enlazables con otros vocabularios, de este modo los agentes pueden interpretar sus semánticas. La siguiente tabla nos presenta los siguientes parámetros: vocabularios propietarios, conjuntos de datos con vocabularios propietarios y nivel de de-referencialidad que puede ser completa, parcial o nula.

Categoría	VP usados	Conjuntos de datos VP	Dereferenciabilidad		
Medios	22 (5,82%)	21 (56,76%)	Completa 0,00 %	Parcial 9,09 %	Ninguna 90, 91 %

Tabla IV. Uso y de-referenciabilidad de los vocabularios empleados

Respecto al último apartado que hace referencia a los metadatos el estudio de los mismos se divide en los siguientes cuatro puntos: suministro de información de procedencia, información de licencias, proporcionar metadatos a nivel de *dataset* o conjuntos de datos y proporcionar metadatos de acceso alternativo.

En cuanto la información de procedencia los datos totales hacen referencia a que el 35,77% de todos los conjuntos de datos utilizan un vocabulario de procedencia. En cuanto a los vocabularios individuales, 28,37% de todos los conjuntos de datos utilizan DC o DCTerm, el 10,77% emplea MetaVocab y el 0,77% usa prv o prov (*Provenance Vocabulary Core Ontology*). En el caso particular de los medios de comunicación 5 conjuntos de datos suministran información sobre la procedencia y el 100% de estos casos emplea como esquema el Dublin Core. Estos conjunto son: *Indymedi, Last.FM RDFization of Events, Artists, and Users; New York Times - Linked Open Data; DBTropes* y *Europeana Linked Open Data*.

Categoría	Cualquier vocabulario de procedencia	Uso DC	Uso MV	Uso prv o prov
medios	5 (13.51%)	100 %	*0.00%*	*0.00%*

Tabla V. Información de procedencia

La siguiente variable hace referencia al suministro de información sobre las licencias de uso de los datos. En total, tan sólo el 7,85% de todos los conjuntos de datos proporcionan información de licencias en RDF. Con respecto a los datos por categorías temáticas, el 5,41 % de los datos de los medios de comunicación proporcionan información sobre las licencias, siendo los datos gubernamentales con el 29,47% los que más referencias realizan sobre este ítem.

El tercer subapartado especifica la provisión de metadatos a nivel de *dataset* o conjuntos de datos mediante la utilización del vocabulario VoID (*Vocabulary of Interlinked Datasets*), ya sea como declaraciones en línea en el conjunto de datos o en un archivo VoID separado.

En total, 140 conjuntos de datos (13,46%) utilizan el vocabulario VoID de los que 48 (4,62%) utilizan un mecanismo de retroenlace, 34 de los cuales enlazan a un archivo VoID recuperable.

Categoría	Total	Enlace	Bien reconocido	En línea
Medios	2 (5,41%)	2.70%	0,00%	2.70%

Tabla VI. Metadatos a nivel de "dataset"

En el caso de los medios de comunicación los casos *Webnmasunotraveler* y *DBTropes* proporcionan este tipo de información.

Finalmente, se localizaron 48 (5,89%) conjuntos de datos que empleaban métodos de acceso alternativos. En total los *endpoints* SPARQL aparecen en un 4,54% de todos los conjuntos de datos mientras que Dumps figura en el 3,8%. En el caso de los medios de comunicación la información que se presenta al respecto es la que aparece en la siguiente tabla.

Categoría	Cualquiera	SPARQL	Dump
Medios	1 (2,70%)	0,00%	2,70%

Tabla VII. Métodos de acceso alternativos

EL CATÁLOGO DE LOS DATOS VINCULADOS

El *Mannheim Linked data Catalog* es un instrumento que proporciona información sobre los conjuntos de datos disponibles en la Web en el momento de la realización del último estudio del estado de la nube desarrollado en agosto de 2014. La creación de este catálogo, así como el diagrama de la nube LOD, ha sido financiado por el proyecto de la Unión Europea *Planet Data*. El contenido de dicha herramienta se ha generado a partir de dos fuentes:

1. En abril de 2014 se realizó un rastreo de la web de LD y se analizaron los conjuntos de datos vinculados hallados que cumplían con las prácticas de LD.
2. La comunidad *Linked data* recoge la meta-información de los conjuntos de datos disponibles en el catálogo *datahub.io.*

Por lo tanto, este catálogo contiene una mezcla de los metadatos proporcionados por la comunidad LD y de los metadatos derivados del rastreo de la web *linked data*.

Al consultar dicho catálogo en la categoría "media" nos da como resultado 26 conjuntos de datos, pero haciendo un análisis detallado existen 4 conjuntos que no son estrictamente *dataset* de medios y que figuran asimismo en la categoría publicaciones. Estos casos son: *Traditional Korean*

Vol. 25 Año 2014
ISNN: 1575-9733

Medicine Ontology, *Europeana Linked Open Data*, *Public Library of Veroia*, CulturaLinkedData. De este modo también se entiende en el propio informe final, ya que como hemos comentando anteriormente el conjunto de datos sobre esta materia disponible en la nube en el momento de hacer el rastreo son 22. El listado de todos los conjuntos de datos de esta categoría aparece recogido en el apéndice que se adjunta al final de este artículo.

Si realizamos una clasificación más específica, tomando como base la definición de la categoría de medios de comunicación anteriormente proporcionada, nos encontramos con los resultados que figuran recogidos en la siguiente tabla. Como podemos observar las subcategorías más predominantes son la de prensa y música con 9 y 8 conjuntos de datos respectivamente. En un número menor de casos le siguen los programas de radio y televisión y las películas.

SUBCATEGORÍA	NÚMERO DE CONJUNTOS
Películas	2
Música	8
Programas de televisión y radio	3
Medios de comunicación impresos	9
TOTAL	**22**

Tabla VIII. Medios por subcategorías

A continuación realizaremos un estudio detallado por cada una de las subcategorías anteriormente definidas.

Películas

Dos son los casos que responden a esta clase: *Linked Movie DataBase* y *Poképédia*. Este último caso resulta curioso puesto que si ya es llamativo la realización de una enciclopedia sobre el "universo Pokémon" mucho más que la misma se presente como un conjunto de datos vinculados. Por su parte, *Linked Movie DataBase* tiene "como objetivo la publicación de la primera base de datos de la web semántica abierta para películas, incluyendo un gran número de interrelaciones de varios conjuntos de datos en la nube de datos abierta y referencias a páginas web relacionadas".

Música

En esta categoría se agrupa todo un conjunto de datos recogidos por DBTune.org. Esta entidad alberga una serie de servidores que proporcionan acceso a datos estructurados relacionados con la música en forma de datos vinculados. De esta institución el catálogo *Mannheim* recoge los siguientes casos: *Jamendo*, Artists*: Last.fm*, *Musicbrainz*, *DBTropes*, John Peel sessions, *Magnatune*.

Por su importancia, nos gustaría destacar en este apartado el caso particular *BBC Music* (figura 4). Este caso es un ejemplo de espacio que entiende la Web como un sistema gestor de contenidos con la información distribuida.

Al realizar una consulta sobre un determinado cantante o grupo musical, en primer lugar se nos presenta distintas pestañas en las que se nos ofrece información sobre: videos, *tracks*, información sobre eventos y enlaces a otras páginas. Estos enlaces nos permiten acceder a otros recursos que ofrecen información musical como son *Musicbraiz* –enciclopedia sobre música– o *Last.fm* –servicio de recomendaciones musicales–, o también podemos tener acceso al espacio que los cantantes tienen en redes sociales como *Myspace* y *Twitter*. Finalmente, nos proporciona información a cantantes y grupos del mismo estilo musical.

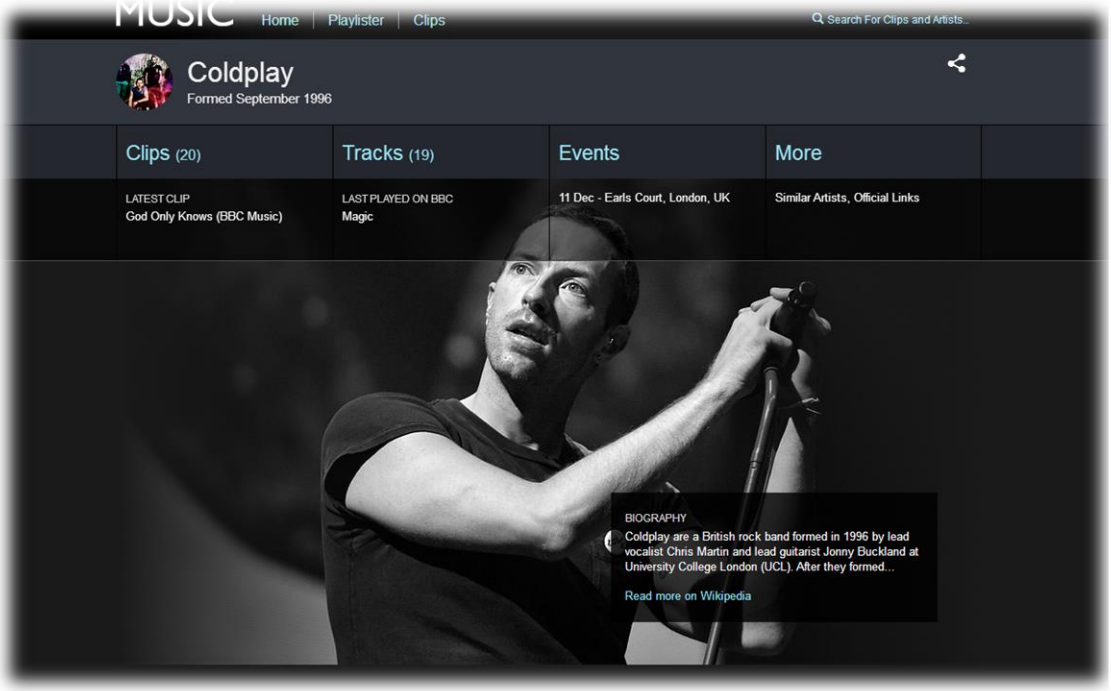

Figura 4. Información sobre *Coldplay* en BBC.Music

Programas de televisión y radio

En esta categoría figuran tres ejemplos: *European Television Heritage*, *BBC Wildlife Finder*, *BBC Programmes*. El proyecto EUscreen está constituido por los archivos de televisión europeos y actúa como un agregador para *Europeana* y la *Biblioteca Digital Europea*. Los otros dos casos pertenecen de nuevo al grupo de comunicación británico BBC. En *BBC Wildlife* (figura 5) podemos encontrar información sobre la fauna, hábitats, adaptaciones y ecozonas. A su vez, dentro de esta última sección se incluyen datos sobre: estado de conservación, descripciones de fondo, fotos, nuevas historias y videos clips de los archivos de la BBC. Por su parte, *BBC Programmes* es el caso más complejo al hacer referencia a los diferentes programas de televisión y radio emitidos por la BBC.

Vol. 25 Año 2014
ISNN: 1575-9733

Figura 5. Nature Wildlife BBC

Medios de comunicación impresos

La última subcategoría pero la más importante en cuanto al número de casos hace referencia a la prensa escrita en donde se incluyen periódicos y revistas. Es así como en este apartado podemos hacer nuevamente una subdivisión entre periódicos, revistas y noticias. Dentro del primer grupo destaca el *New York Times-LOD* (figura 6). Este mítico diario manifiesta en su web de LOD que "durante los últimos 150 años, *The New York Times* ha mantenido uno de los vocabularios de noticias de mayor autoridad jamás desarrollados. En 2009, comenzamos a publicar este vocabulario como datos abiertos enlazados". Sobre la tecnología de datos vinculados, el periódico expresa que utiliza aproximadamente 30.000 etiquetas para enriquecer las páginas con los temas de actualidad. Su objetivo es publicar todas estas etiquetas como datos abiertos enlazados.

Vol. 25 Año 2014
ISNN: 1575-9733

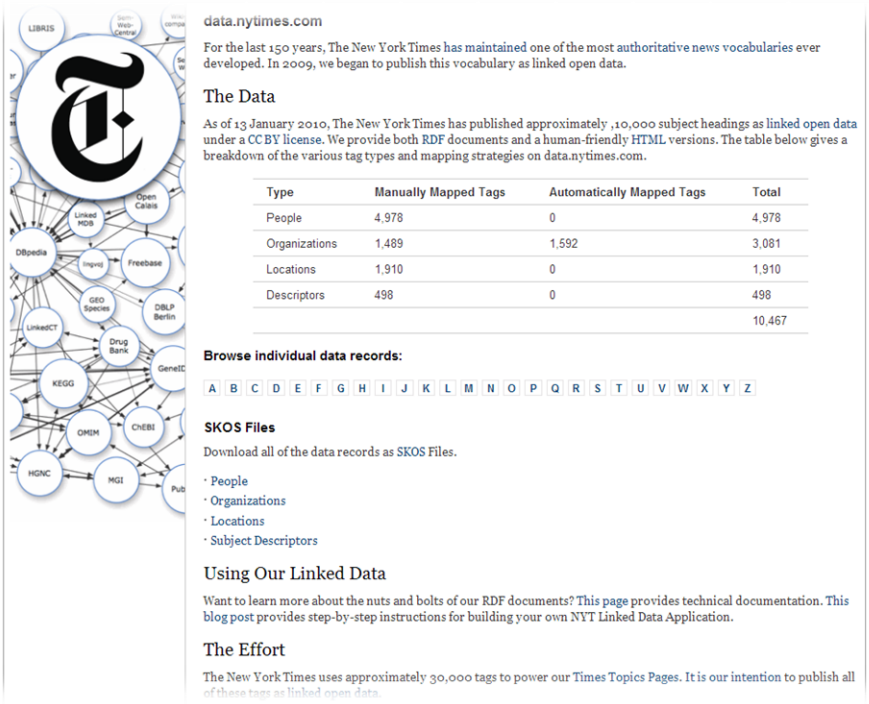

Figura 6. The New York Time-LOD

En el caso de la prensa escrita nos gustaría destacar también la única iniciativa española que se encuentra disponible, en estos momentos, en la nube de los datos de los medios de comunicación. En concreto se trata del caso denominado *Webnmasunotraveler*. Este proyecto incluye periódicos y plataformas digitales del grupo Prisa como son: *Suplemento El País*, *Guías Aguilar*, *Canal Viajar* y *Prisa Digital*. Asimismo, también se incluyen las recomendaciones de los usuarios, sus fotos y blogs en los que se relatan experiencias sobre viajes realizados por todo el mundo.

En el apartado de los casos que recogen artículos y noticias podemos citar en primer lugar por su importancia *Ontos News Portal*. Este *dataset* extrae "hechos" (objetos, como personas u organizaciones, así como las relaciones entre ellos, por ejemplo, una persona está trabajando para una organización o vive en un lugar). Los hechos se fusionan juntos y construyen una enorme red de información que incluye referencias a los artículos respectivos.

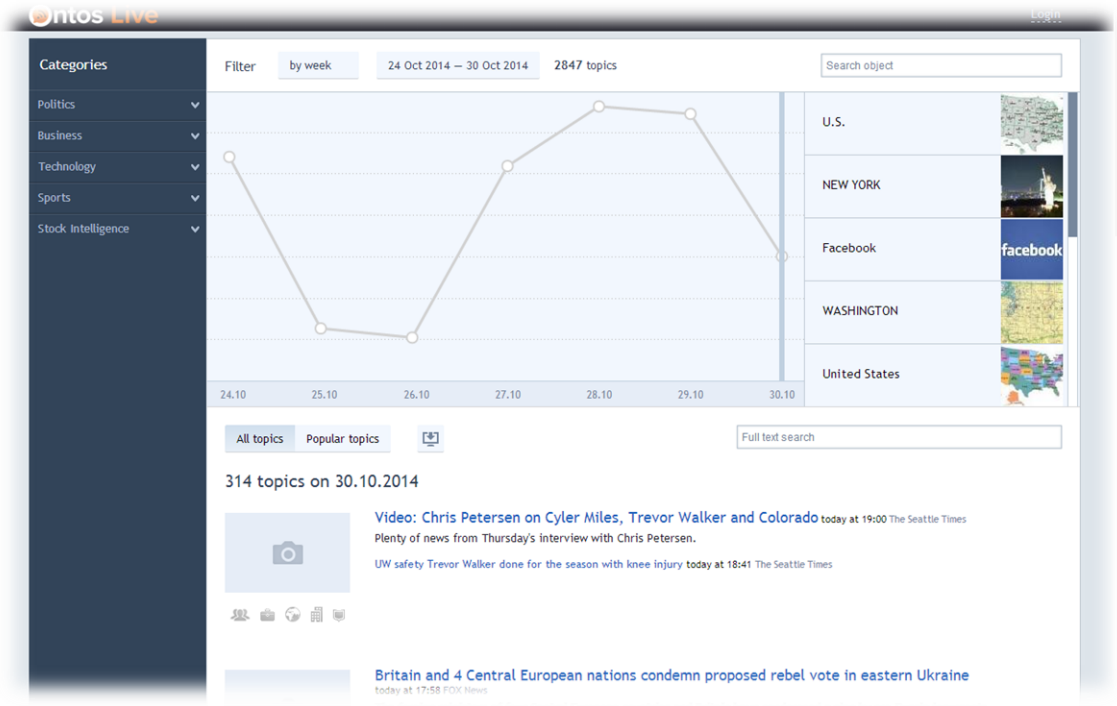

Figura 7. Ontos News Portal

También es interesante el caso *IPTC*, el cual no sólo proporciona formatos de intercambio de noticias sino que también crea y mantiene conjuntos de conceptos que pueden asignarse como valores de metadatos para objetos de noticias: texto, fotografías, gráficos, archivos de audio y vídeo y *streams*. Este hecho permite una codificación consistente de metadatos de noticias a lo largo del tiempo.

Otros conjuntos de datos de este subgrupo serían: *Indymedia* y *OpenCalais*. El primero se trata de una red global participativa de periodistas independientes que informan sobre temas políticos y sociales; el segundo, es un servicio que extrae metadatos semánticos a partir del contenido textual de las páginas web. El único caso correspondiente a una revista es el denominado *Bombsite* que hace referencia a *Boom Magazine* pero sobre este caso apenas se proporciona información.

Finalmente, debemos hacer referencia al caso denominado *Public Record Office Victoria Semantic Wiki*. Tal y como figura en la propia descripción del caso, esta oficina es el archivo del Gobierno del Estado de Victoria, Australia. La colección incluye recuerdos de acontecimientos y decisiones que han dado forma a la historia de la Colonia y del Estado de Victoria, así como los registros de inmigración y el transporte marítimo, los juicios penales y prisiones, primeros ministros y gobernadores, comisiones reales, las juntas de investigación, los testamentos y validaciones, entre otros muchos documentos. Por la propia definición del caso consideramos que hubiera sido más apropiado incluir este conjunto dentro de la categoría denominada “publicaciones”, junto con el

resto de conjuntos pertenecientes a archivos, bibliotecas, galerías y museos, que etiquetarlo como un caso relativo a los medios de comunicación.

CONCLUSIONES

Si "la digitalización ha supuesto la mayor evolución tecnológica para los medios de comunicación en toda su historia" (Caldera y Arranz, 2012, p. 31) la aparición de los movimientos *data, open, linked,* supondrá una nueva revolución en el modo de concebir los medios informativos. Es así como la tecnología data, más concretamente *big data*, cambiará los modos de producción y difusión de la información. Por ejemplo, en el caso de la prensa, el mismo diario será diferente en función de los intereses personales de cada lector, y así, en un caso se destacarán las noticias económicas y políticas, y en otro, las culturales y deportivas tras el análisis individual de las preferencias de cada usuario.

De la web de los documentos hemos pasado a la web de los datos en la que los diferentes datos se conectan y vinculan entre sí a través de una serie de pautas perfectamente establecidas. Los datos así publicados se recogen en la denominada "nube de datos".

En el presente artículo hemos intentando realizar una radiografía del estado de los datos vinculados en el caso particular de los medios de comunicación. Realizando un resumen global de los principales resultados obtenidos podemos decir que: es muy positivo que los datos de los *media* estén representados en la nube, sin embargo, el número de casos es muy reducido si se compara con otros grupos, como pueden ser el de la web social o los datos gubernamentales. En lo referente a la interconexión, la mitad de los conjuntos enlazan a otros siendo el predicado más utilizado en la interconexión "owl:sameAs". Existe un ligero predominio de los vocabularios propietarios frente a los no propietarios destacando la hegemonía del uso de Dublin Core como vocabulario de procedencia. Por otro lado, apenas se suministra información sobre la licencia de uso de los datos y en lo referente a la provisión de datos a nivel de *dataset* prevalece el uso de VoID frente a otros métodos de acceso alternativos. La presencia de estas últimas variables en el caso de los medios de comunicación son muy minoritarias frente a los otros grupos presentes en la nube.

Tras el análisis particular de los conjuntos de datos referentes a los medios, podemos decir, que existen casos muy representativos de datos vinculados, como puede ser el caso de *BBC music*, dentro de la categoría de información musical o el *New York Times-LOD* en el caso de los medios impresos. Con el examen detallado de cada ejemplo hemos intentado conseguir un doble objetivo: ilustrar con ejemplos el fenómeno de LOD en los medios y ayudar a esclarecer los pasos a seguir para publicar la información como datos abiertos y vinculados.

Para terminar nos gustaría resaltar posibles líneas de investigación dentro de este ámbito. Por un lado, se debería profundizar en el estudio pormenorizado de cada caso individual. Asimismo, debería analizarse como los casos particulares de los *dataset* de los medios de comunicación se

enlazan entre sí. El siguiente paso sería ver como estos *dataset* se vinculan con los conjuntos de datos de las otras categorías presentes en la nube. En este sentido podría estudiarse la relación con el caso específico de los datos GLAM. Sería interesante, por ejemplo, ver si las noticias se etiquetan empleando los vocabularios controlados pertenecientes a las grandes instituciones bibliotecarias o se ilustran con fondos procedentes de importantes galerías y museos.

REFERENCIAS BIBLIOGRÁFICAS

Bennett, Daniel; Harvey, Adam (2009). Publishing Open Government Data. [S.l] : W3C, 2009 http://www.w3.org/TR/gov-data (2014-11-11)

Bernes-Lee, Tim (2006). Linked data – Design Issues. [S.l.] : Bernes-Lee, 2006. http://www.w3.org/DesignIssues/LinkedData.html (2014-11-11)

Bernes-Lee, Tim (2010). Is your Linked Open Data 5 Star?. [S.l.] : Bernes-Lee, 2010. http://www.w3.org/DesignIssues/LinkedData.html (2014-11-11)

Budapest Open Access initiative (2002). Budapest: BOAI, 2002 http://www.opensocietyfoundations.org/openaccess (2014-11-11)

Caldera Serrano, Jorge ; Arranz Escacha, Pilar (2012). Documentación audiovisual en televisión. Barcelona : UOC, 2012. ISBN 978-84-9029-982-1

DataHub (2014). Cambridge: Open knowledge, 2014 http://datahub.io/de/group/lodcloud (2014-11-11)

Europeana Linked Open Data (2014). European Union: Europeana Fundation: Europenan Creative Project, 2014. http://data.europeana.eu (2014-11-11)

Ferrer, Antonia; Peset, Fernanda; Aleixandre, Rafael. (2011). Acceso a los datos públicos y su reutilización: open data y open government. // El profesional de la información. ISSN: 1699-2407. 20:3 (mayo-junio 2011) 260-269. http://elprofesionaldelainformacion.metapress.com/media/6n99qgwhmg2rlu7pnqt1/contributions/9/2/7/4/92741636q145x727.pdf (2014-11-11) DOI: 10.3145/epi.2011.may.??

Ferreras, Tránsito. (2011). Open Access en España: los Repositorios Institucionales, 2011. // V Jornadas de e-learning en la formación para el empleo en las Administraciones Públicas, Valladolid, 15-17 de septiembre de 2011. http://www.slideshare.net/Transito09/open-access-en-espaa-los-repositorios-institucionales (2014-11-11)

Gray, Jonathan; Bounegru, Liliana; Chambers, Lucy (eds.) (2012). Data journalism handbook: how journalist can use data to improve the news. Maastricht: European Journalism Center; Cambridge: Open knowledge, 2012. http://datajournalismhandbook.org/1.0/en/index.html

Hassanzadeh, Oktie; Consens, Mariano P. (2008) Linked Movie DataBase.[S.l.]:[s. n.], 2008. http://linkedmdb.org/(2014-11-11)

Heath, T., & Bizer, C. (2011). Linked data: Evolving the Web into a Global Data Space. Florida, USA: Morgan & Claypool Publishers, 2011.

Hernández, Tony; García, María Antonia (2013). Datos abiertos y repositorios de datos: nuevo reto para los bibliotecarios. // El profesional de la información. ISSN: 1699-2407. 22 (3) (mayo-junio 2013) 259-263.

http://www.thinkepi.net/datos-abiertos-repositorios-datos-nuevo-reto-bibliotecarios DOI: http://dx.doi.org/10.3145/epi.2013.may.10

Isaac, Antoine; Waites, William; Young, Jeff; Zeng, Marcia. Library Linked data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets. [S.l]: W3C, 2011. http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025 (2014-11-11)

Linked data : connect Distributed Data across the Web (2014). [S.l.]: [s.n.], 2014. http://linkeddata.org (2014-11-11)

LinkingOpenData (2014). [S.l.]: W3C, 2014. http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData (2014-11-11)

The Linking Open Data cloud diagram (2014). [S.l.] : Linking Open Data, 2014 http://lod-cloud.net/ (2014-11-11)

Mannheim Linked data Catalog (2014). Mannheim: University of Mannhein, 2014. http://linkeddatacatalog.dws.informatik.uni-mannheim.de (2014-11-11)

Martín, Sandra ; Angelozzi, Silvina. (2013). Datos abiertos enlazados y libros abiertos: impacto en las bibliotecas y en el desarrollo de la sociedad de la información. // 42 JAIIO : Jornadas Argentinas de Informática, SSI 2013: 11º Simposio sobre la Sociedad de la Información, Cordoba, Argentina. http://hdl.handle.net/10760/20153 (2014-11-11) (2014-11-11)

Mazzo Iturriaga, Rodrigo (2010). Linked Open Data: qué es y ejemplos en el mundo. Chile : Biblioteca del Congreso Nacional de Chile, 2010. http://www.bcn.cl/de-que-se-habla/open-data-link-data (2014-11-11)

Mitchel, Erik (2013). Library Linked data: research and adoption. // Library Technology Reports. 5:49 (2013) 11-25.

Open definition (2014). Cambridge: Open knowledge, 2014. http://opendefinition.org/ (2014-11-11)

Open Source Initiative (1998). California: OSI, 1998. http://opensource.org/osd (2014-11-11)

Peset, Fernanda; Ferrer, Antonia; Subirats, Inma. (2011). Open data y Linked open data: su impacto en el área de bibliotecas y documentación. // El profesional de la información. ISSN: 1699-2407. 20:2 (marzo-abril 2011) 165-173. DOI: 10.3145/epi.2011.mar.06

Ríos, Ana B. Linked Open Data. // DINLE: Diccionario Digital de Nuevas Formas de Lectura y Escritura. Salamanca: Ediciones Universidad de Salamanca; [S.l.]: Red Internacional de Universidades Lectoras, 2014. http://dinle.eusal.es/searchword.php?valor=Linked+Open+Data (2014-11-11)

Ríos, Ana B.; Ferreras, Tránsito; Martín, Diego. From Bibliographic Records to Data: changes in the library environment with the application of Linked Open Data technologies // Information Resources Management Journal. ISSN: 1040-1628. 27:3 (July-September 2014) 28-41.

Schmachtenberg, Max; Bizer, Christian; Paulheim, Heiko (2014) . Adoption of the linked data best practices in diferent topical domains. Mannheim: University of Mannhein; Unión Europea: Planet Data, 2014. http://dws.informatik.uni-mannheim.de/fileadmin/lehrstuehle/ki/pub/SchmachtenbergBizerPaulheim-AdoptionOfLinkedDataBestPractices.pdf (2014-11-11)

Schmachtenberg, Max; Bizer, Christian; Paulheim, Heiko (2014). State of the LOD Cloud 2014. Mannheim: University of Mannhein; Unión Europea: Planet Data, 2014. http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/#toc1 (2014-11-11)

APÉNDICE. Relación de casos en la categoría medios de comunicación

CASOS	ENLACE
PELÍCULAS	
Linked Movie DataBase	http://linkedmdb.org/
Poképédia	http://www.pokepedia.fr/
MÚSICA	
DBTune.org Jamendo RDF Server	http://dbtune.org/jamendo/
DBTune.org Artists: Last.fm	http://dbtune.org/artists/last-fm/
DBTune.org Musicbrainz D2R Server	http://dbtune.org/musicbrainz/
Last.FM RDFization of Events, Artists, and Users	http://lastfm.rdfize.com/
DBTropes	http://skipforward.opendfki.de/wiki/DBTropes
DBTune.org John Peel sessions RDF server	http://dbtune.org/bbc/peel/
BBC Music	http://www.bbc.co.uk/music
DBTune.org Magnatune RDF server	http://dbtune.org/magnatune/
TELEVISIÓN Y RADIO	
European Television Heritage	http://lod.euscreen.eu/
BBC Wildlife Finder	http://www.bbc.co.uk/nature/wildlife
BBC Programmes	http://www.bbc.co.uk/programmes
PRENSA	
PERIÓDICOS	
Webnmasunotraveler	http://webenemasuno.linkeddata.es/

Vol. 25 Año 2014
ISNN: 1575-9733

New York Times - Linked Open Data	http://data.nytimes.com/
Chronicling America	http://chroniclingamerica.loc.gov/about/api/
	REVISTAS
Bombsite	http://bombmagazine.org/
	ARTÍCULOS Y NOTICIAS
Public Record Office Victoria Semantic Wiki	http://www.wiki.prov.vic.gov.au/index.php/Public_Record_Office_Victoria_Semantic_Wiki
Ontos News Portal	http://news.ontos.com/
Indymedia	http://www.indymedia.org/es/
OpenCalais	http://www.opencalais.com/
IPTC NewsCodes	http://www.iptc.org/cms/site/index.html?channel=CH0103