Enlazando datos, descubriendo objetos: Aplicación de las tecnologías LOD en las instituciones patrimoniales

Ana B. Ríos Hilario

Departamento de Biblioteconomía y Documentación, Univerdiad de Salamanca, España, anarihi@usal.es

Resumen

Se analiza la representación de los conjuntos de datos pertenecientes a Galerías, bibliotecas, archivos y museos (GLAM) en la nube de los datos abiertos y vinculados. En primer lugar, se establecen las bases teóricas en las que se sustenta el análisis: por un lado, se define el concepto de linked open data (LOD); posteriormente, se delimitan las fuentes objeto de estudio, en concreto: el diagrama de la nube LOD en el que se exponen la totalidad de los conjuntos de datos; el Catálogo de datos enlazados, en donde se obtiene información individualizada de cada caso específico y el informe titulado Conjunto de datos, vocabularios de valores y conjuntos de elementos de metadatos, cuya tipología nos sirve para categorizar los datasets bibliotecarios. El análisis propiamente dicho se centra: primero en el estudio de la muestra de los datos examinados sobre la totalidad de la nube; segundo, en la categorización de dichos de conjuntos en función de la clasificación previamente establecida; finalmente, se describen los casos más representativos centrándose en la tecnología de datos vinculados empleada. Se concluye exponiendo los principales resultados obtenidos y estableciendo una serie de propuestas de mejora.

Palabras claves: *Datos abiertos y vinculados (LOD). Galerías, bibliotecas, archivos y museos (GLAM). Diagrama de la nube LOD. Catálogo de datos enlazados Mannheim. Conjunto de datos, vocabularios de valores y conjuntos de elementos de metadatos (VOCABDATASET).*

Abstract

The representation of the data sets belonging to Galleries, libraries, archives and museums (GLAM) cloud of linked open data is analyzed. First, theoretical basis on which the analysis is based are set: on one hand, the concept of open data liked (LOD) is defined; subsequently object of study sources are delimited, namely: the cloud diagram LOD in which all datasets are displayed; Mannheim Linked Data Catalog where individual information of each specific case is obtained and the report entitled Datasets, Value Vocabularies, and Metadata Element Set whose typology serves to categorize librarians datasets. The analysis itself focuses: first study sample of data examined of the entire cloud; second, in the categorization of such set according to the predetermined classification; finally, the most representative cases focusing on technology linked data used are described. It is concluded stating the main results and establishing a series of proposals for improvement.

Keywords: *Linked Open Data (LOD). Galleries, libraries, archives and museum (GLAM). Cloud diagram LOD. Mannheim Linked Data Catalog. Datasets, Value Vocabularies, and Metadata Element Set (VOCABDATASET).*

1. Introducción

La web semántica es una web extendida, dotada de mayor significado en la que cualquier usuario en internet podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida. Al dotar a la web de más significado y, por lo tanto, de más semántica, 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 (*Guía breve de Web Semántica*, 2017). Es así como podemos hablar de 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*, 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).*

Desde hace más de dos décadas, en el entorno de las instituciones patrimoniales o de las entidades también conocidas con las siglas de LAM o GLAM (*galleries, libraries, archives* and *museum*) se están produciendo toda una serie de cambios que han tenido como consecuencia un replanteamiento total en los modos de gestión y funcionamiento de dichas instituciones.

Tomando como base las anteriores premisas el objetivo principal de este artículo se centra en estudiar los conjuntos de datos abiertos y vinculados pertenecientes a las instituciones patrimoniales. De este propósito general se establecen a su vez los siguientes objetivos específicos:

- Definir los conceptos claves objeto de estudio en lo referente a los datos abiertos y vinculados y delimitar las fuentes de información empleadas específicamente para la realización del análisis.
- Establecer una categorización de los distintos *datasets* GLAM presentes en el *LOD cloud diagram (*diagrama de la nube LOD*)* siguiendo la tipología establecida en el documento *Datasets, Value Vocabularies, and Metadata Element Sets* (*Conjunto de datos, vocabularios de valores y conjuntos de elementos de metadatos*; VOCABDATASET) realizado por el World Wide Web Consortium (W3C).
- Realizar un análisis estadístico que posibilite conocer la representación de los conjuntos bibliotecarios dentro de la nube en general y de modo más específico la representación de las categorías establecidas.
- Detallar los casos más representativos de cada categoría expuestos en el *Mannheim Linked Data Catalog* (*Catálogo de datos enlazados Mannheim*) centrándonos especialmente en la tecnología de datos vinculados empleada.

Dejando constancia de la imposibilidad de recoger y analizar la totalidad de *datasets* del ámbito GLAM publicados como datos vinculados en la actualidad, decidimos tomar como fuente de información principal el *LOD cloud diagram* debido a su representatividad y vigencia y cuya última actualización se ha realizado en febrero de 2017. Sin embargo, para obtener datos estadísticos específicos de cada conjunto de datos tenemos que acudir al rastreo de la web *linked data* que se efectuó en abril de 2014 y su difusión se hizo pública en agosto de ese mismo año. Por lo tanto, la ausencia de documentación detallada sobre cada *datasets* en la última versión de la nube hace que los datos que figuren a continuación sean aproximados. No obstante, el objetivo último de nuestro texto no es tanto la cuantificación de los *datasets* sino la categorización de los mismos que es precisamente donde reside la validez de nuestro análisis.

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 different topical domains* (Schmachtenberg; Bizer y Paulheim, 2014a) y su versión

abreviada disponible en el recurso *State of the LOD cloud* 2014 (Schmachtenberg; Bizer y Paulheim, 2014b). Así mismo, para el análisis específico de los *datasets* acudimos al *Mannheim Linked Data Catalog.* Otro instrumento clave para el desarrollo del estudio ha sido el *Datasets, Value Vocabularies, and Metadata Element Sets* (Isaac; Waites; Young; Zeng, 2011) fruto del informe final realizado por el *Library Linked Data Incubator Group (Grupo Incubador de Datos Vinculados de Bibliotecas*).

Para cumplir con los objetivos propuestos hemos aplicado una triple metodología. En primer lugar, para sentar las bases teóricas se acudió a las propias fuentes anteriormente indicadas. Para la cuantificación de los *datasets* se llevó a cabo un análisis estadístico y, finalmente, se aplicó el método descriptivo para definir los casos más representativos.

El artículo se estructura en dos partes. En la primera se define el concepto de *linked open data* y se explican las fuentes de información en las que se basa el análisis. En la segunda se presenta el análisis propiamente dicho: tras la categorización y contabilidad de los conjuntos de datos bibliotecarios se muestran los casos más relevantes. Finalmente, en las conclusiones se exponen los principales resultados obtenidos y se realizan una serie de propuestas de mejora. El artículo incluye un apéndice final en el que figuran categorizados según las clases establecidas todos los conjuntos de datos bibliotecarios.

2. Conceptos claves entorno a la definición de linked open data (LOD)

2.1. Linked Open Data (LOD)

Ultimamente, la palabra "data" aparece con frecuencia en la literatura científica, generalmente acompañada de otro término que delimita su significado como puede ser *open data*, *data mining* y el más reciente, *big data*. 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).

Por supuesto, en este laberinto de datos también están presentes los datos bibliotecarios cuya misión se centra en el potencial para la creación de estos datos globalmente interconectados; el intercambio y utilización conjunta de datos con instituciones no bibliotecarias; la creciente confianza en el crecimiento de la web semántica, y en el mantenimiento de una gráfica cultural global de la información que sea fiable y persistente (Bauer y Kaltenböck, 2012). Como ejemplos más representativos de datos bibliotecarios podemos citar la presencia de bibliotecas como la Library of Congress y de iniciativas como Europeana y la Digital Public Library of America (DPLA).

El concepto de *open* se incluye dentro de una expresión más amplia que es la de conocimiento abierto. Según el Open Definition Advisory Council (2014) este término 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".

Los datos enlazados es la forma que tiene la web semántica de vincular los distintos datos que están distribuidos en la

Web, de tal manera que los datos se enlazan del mismo modo que lo hacen las páginas web. Si queremos una definición precisa del término podemos acudir a la página oficial *Linked data* 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".

Para hablar de *linked data* (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.

2.2. La nube de los datos abiertos y enlazados

El *LOD cloud diagram* (Diagrama de la nube LOD) expone 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 (Figura 1). El propósito principal de este proyecto es "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 estable-cimiento de enlaces RDF entre elementos de datos de diferentes fuentes" (Linking Open Data, 2017). El *Data Hub* a su vez podemos definirlo 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 y Zeng, 2011). Al hacer clic en la imagen nos lleva a una versión SVG interactiva, donde cada conjunto de datos es un hipervínculo a su entrada en el DataHub.

La interpretación del diagrama es sencilla. La figura 1 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 corresponde al número de aristas conectadas a cada conjunto de datos. Los números se calculan basándose en conjuntos de datos conectados en el diagrama. La línea indican la existencia de al menos un enlace entre dos conjuntos de datos. Un enlace, para nuestros propósitos, es una tripleta RDF en la que sujeto y objeto URIs están en los espacios de nombres de diferentes conjuntos de datos. En la versión interactiva, el color de la línea indica la dirección del enlace, por ejemplo, si un enlace de A a B es de color verde entonces significa que el conjunto de datos A contiene tripletas RDF que utiliza identificadores de B, y si se lee al revés, significa que ese conjunto de datos B contiene tripletas RDF que emplea identificadores de A.

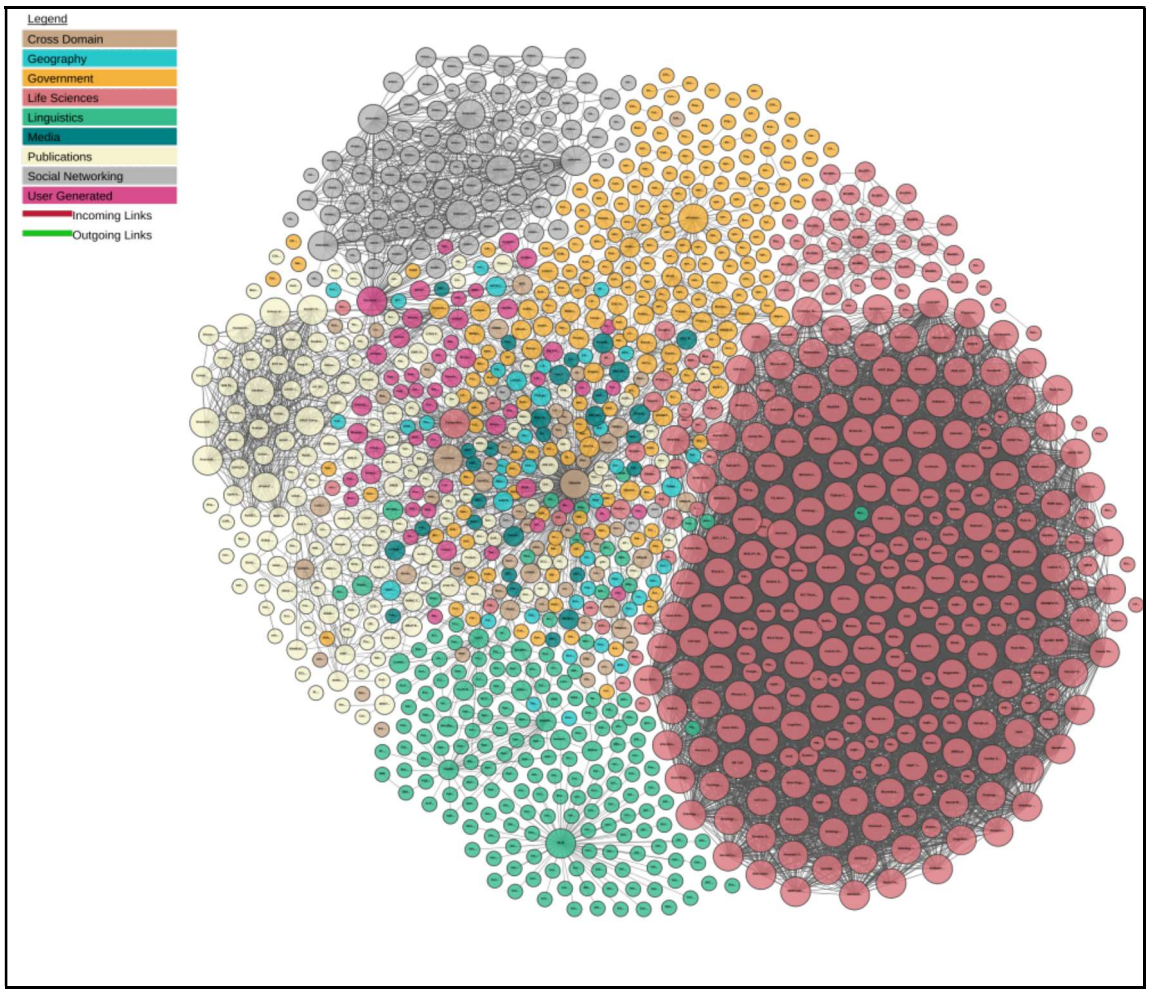

Figura 1. Diagrama de la nube LOD (Fuente: http://lod-cloud.net/)

Los conjuntos de datos referentes a la categoría *Publicaciones*, donde se incluyen los datos bibliotecarios en su más amplio sentido, se identifican por el color márfil y se sitúan en el margen izquierdo del diagrama. Ha habido un cambio significativo respecto a las anteriores versiones en donde este conjunto de datos estaban coloreados en verde claro y se posicionaba en el margen derecho del diagrama.

2.2.1. El catálogo de datos enlazados

El *Mannheim Linked Data Catalog* (*Catálogo de datos enlazados Mannheim*) 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, han sido financiados 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 metainformación de los conjuntos de datos disponibles en el catálogo datahub.io.

Dicho catálogo es un complemento al documento *Adoption of the linked data best practices in different topical domains* (Schmachtenberg; Bizer; Paulheim, 2014a) "permitiendo al lector desglosar y explorar la información sobre los conjuntos de datos detrás de cada resultado estadístico" (Schmachtenberg; Bizer; Paulheim, 2014b). Precisamente, en este trabajo nuestro objetivo es el estudio de los *datasets* bibliotecarios etiquetados bajo la categoría de *Publicaciones* y que figuran en la mencionada herramienta. Sin embargo, y tal y como se explica en la propia web de la nube, esta documentación no se ha realizado sobre la última actualización del diagrama.

2.3. La nube de los datos abiertos y enlazados

El W3C Library Linked Data Incubator Group (Grupo Incubador del W3C de Datos Vinculados de Bibliotecas; LLD-XG) se crea en el año 2010 con el objetivo de "contribuir a incrementar la interoperabilidad global de los datos de las bibliotecas en la Web, reuniendo a personas implicadas en actividades de la web semántica —centradas en los datos vinculados— en bibliotecas e instituciones afines, mediante el examen de las iniciativas en curso e identificando futuras vías de colaboración". Fruto de su trabajo durante un año fue la redacción de un informe final que se publica en octubre de 2011 bajo la denominación de Library Linked Data Incubator Group Final Report (Informe Final del Grupo Incubador de Datos Vinculados de Bibliotecas).

Dicho informe se estructura en cuatro apartados: ámbito del informe; beneficios; estado de la cuestión y recomendaciones.

El informe final incluye a su vez dos documentos que lo complementan pero que se publican de modo independiente: *Use Cases* (Casos de usos; USECASE) y *Datasets, Value Vocabularies, and Metadata Element Sets* (Conjunto de datos, vocabularios de valores y conjuntos de elementos de metadatos; VOCABDATASET).

El USECASE presenta los casos de uso recopilados y ofrece un breve resumen de cada caso individual. Sin embargo, para el propósito de nuestro estudio el documento más relevante es el titulado *Conjunto de datos, vocabularios de valores y conjuntos de elementos de metadatos*. Su objetivo es la identificación de "un conjunto de recursos útiles para la creación o consumo de datos enlazados en el campo de las bibliotecas" (Isaac; Waites; Young y Zeng, 2011). La lectura de este informe permitirá, por un lado, "proporcionar a la comunidad de datos enlazados la oportunidad de comprender el punto de vista específico, los recursos y la terminología usada por parte del colectivo bibliotecario para sus datos", al mismo tiempo, "servirá de ayuda a la los profesionales de la información a comprender el modo en el que los conceptos relacionados con los datos enlazados encajan con sus propias tradiciones profesionales" (Isaac; Waites; Young y Zeng, 2011). Se concibe como un punto de partida para que los profesionales puedan encontrar, comprender y explorar algunos

ejemplos de conjuntos de datos, vocabularios controlados y conjunto de metadatos. El informe comienza por definir los conceptos claves que dan nombre al propio documento. Posteriormente de cada uno de ellos realiza un listado de los casos de uso recopilados en el documento precedente.

Ya que estos mismos términos son los que vamos a emplear a la hora de categorizar los datasets bibliotecarios recogidos en el catálogo *Mannheim*, a continuación pasaremos a definirlos según se detalla en el propio VOCABDATA-SET (Isaac; Waites; Young y Zeng, 2011).

- Conjunto de datos: considerados como colecciones de metadatos estructurados, descripciones de cosas como los libros de una biblioteca. Los registros bibliográficos son afirmaciones sobre cosas, en los que cada afirmación está formada por un elemento ("atributo" o "propiedad") de la entidad y un "valor" para ese elemento.

  - Ejemplo: un registro de un conjunto de datos sobre un libro determinado puede tener un elemento 'materia' tomado de *Dublin Core* y un valor para la materia tomado de *Library of Congress Subject Headings* (LCSH).

- Vocabulario de valores: un vocabulario de valores define recursos (como instancias de materias, estilos artísticos o autores) que se utilizan como valores de elementos en los registros de metadatos. Así, un vocabulario de valores es una lista de los valores permitidos para un elemento. Ejemplos de ello son: tesauros, listas de códigos, listas de términos, esquemas de clasificación, listas de encabezamientos de materia, taxonomías, ficheros de autoridades, nomenclátores geográficos, esquemas de conceptos y otros tipos de sistemas de organización del conocimiento.

  - Ejemplo: el Fichero Virtual de Autoridades (Virtual International Authority File; VIAF) define autoridades de nombres (p.e, Mark Twain).

- Conjuntos de elementos de metadatos o conjuntos de elementos: los conjuntos de elementos de metadatos definen las clases y atributos utilizados para describir entidades de interés. En la terminología de datos vinculados, estos conjuntos de elementos se concretan generalmente por medio de RDF Schemas (esquemas RDF), Web Ontology Language (Lenguaje de Ontologías Web; OWL), que con frecuencia se agrupan bajo el término "vocabulario RDF". Normalmente los conjuntos de elementos de metadatos no describen entidades bibliográficas, sino más bien pro-porcionan los elementos que se pueden utilizar para describir estas entidades. Términos equivalentes serían: vocabula-rios RDF, esquemas RDF y ontologías (*Library terminology informally explained*, 2011).

  - Ejemplo: MARC21 define los elementos (campos) para describir registros bibliográficos y de autoridades.

3. Los datos bibliotecarios en el LOD cloud diagram

3.1. Análisis estadístico

Antes de comenzar el estudio pormenorizado de la categoría denominada *Publicaciones* es necesario obtener una visión global de la totalidad de conjuntos de datos presentes en el diagrama (Tabla 1). Como ya hemos comentado anteriormente, en esta última actualización no se ha llevado a cabo un estudio deslindado de cada uno de los *datasets*. Es así como frente a los 1014 *datasets* del año 2014, en la actualidad en la nube se recogen un total de 1139 conjuntos. Mediante esta anotación podremos obtener una visión aproximada de la representación de los datos patrimoniales en la actualidad.

Tabla 1. Conjuntos de datos por categorías temáticas

Categoría	Conjunto de datos	Porcentaje
Gobierno	183	18,05%
Publicaciones	96	9,47%
Ciencias de la vida	83	8,19%
Usuario	48	4,73%
Dominios cruzados	41	4,04%
Medios	22	2,17%
Geografía	21	2,07%
Web social	520	51,28%
Total	1014	100%

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

Como puede observarse de las 8 materias identificadas, la categoría *Publicaciones* ocupa la tercera posición, con 96 conjuntos de datos identificados, lo que se corresponde con el 9,47% del espacio en la nube. Destaca por encima del resto, los conjuntos de datos relativos a la web social (520) que ostenta la primera posición.

Centrándose en el análisis especifico de la materia objeto de estudio, en primer lugar, tenemos que comenzar definiendo que *datasets* se han etiquetado bajo dicha denominación.

Tal y como se indica en el propio documento la categoría denominada *Publicaciones* comprende "conjuntos de datos de bibliotecas, información sobre publicaciones científicas y conferencias, listas de lectura de las universidades, y las bases de datos de citas".

El número de conjuntos que figura en el resultado final difiere del que se muestra en el catálogo *Mannheim*. En dicho catálogo aparecen 101 conjuntos de datos etiquetados dentro de la categoría *Publicaciones*. Tal diferencia viene dada porque los conjuntos pueden categorizarse en ocasiones en más de una categoría pero a la hora de realizar el recuento final se ha optado por incluirlos en una única. De tal manera que del total de casos que figuran en el catálogo, un ejemplo

aparece repetido y cuatro son contabilizados además en otra categoría. Es así como Uniprot-citation se incluye en el apartado de Ciencias de la vida, y DWS-Group, Data.dcs y Morelab se recogen en la tipología denominada Web Social.

Si hacemos una clasificación del total de los 96 casos contabilizados atendiendo a las subcategorías establecidas en la anterior definición obtenemos los siguientes resultados (Tabla 2).

Tabla 2. Subcategorías de la categoría Publicaciones

Subcategoría	Número	Porcentaje
Bibliotecas (GLAM)	42	44%
Publicaciones científicas	23	24%
Listas de lectura	17	18%
Bases de datos	14	14%
Total	96	100%

Como puede observarse el mayor número de casos hace referencia a la subcategoría denominada *Bibliotecas* con un total de 42 ejemplos. Entiéndase el término bibliotecas en su más amplio sentido, es decir, abarcaría a todas las instituciones recogidas bajo las siglas GLAM. En un porcentaje similar se encuentran tanto la subcategoría que recoge las *Publicaciones científicas y conferencias* como las *Listas de lectura* ofertadas por las instituciones académicas de enseñanza superior. La última subcategoría referente a las *Base de datos de citas* presentan 15 casos recogidos en el mencionado catálogo.

Atendiendo a la clasificación que recoge el documento VOCABDATASET que divide los datos vinculados bibliotecarios entre: conjuntos de datos, vocabulario de valores y conjunto de elementos de metadatos, de los 42 conjuntos de datos obtenidos en la anterior tabla como *datasets* bibliotecarios podemos obtener los siguientes datos:

Tabla 3. Clases según la clasificación establecida por el VOCABDATASET

Clases	Número	Porcentaje
Colecciones	20	48
Vocabularios de valores	16	38
Metadatos	6	14
Total	42	100

Las *colecciones* bibliotecarias destacan ligeramente sobre los *vocabularios de valores*. En un porcentaje menor aparecen los conjuntos de *metadatos*.

Haciendo nuevamente un desglose del conjunto de casos que se han agrupado bajo la clase denominada *Colecciones*, aparece en porcentajes similares tanto los fondos de bibliotecas y archivos, bibliografías nacionales y catálogos (Tabla 4). Aunque todos y cada uno de los casos referentes a esta categoría hacen referencia a "los conjuntos de datos considerados como colecciones de metadatos estructurados" (VOCABDATASET, 2011) los hemos diferenciado entre las

diferentes colecciones que figuran en la tabla 3, tan sólo hemos dejado un caso de catalogar por su dificultad de clasificar en dichas categorías. En concreto dicho caso hace referencia al Project Gutenberg que consta de un conjunto de libros electrónicos gratuitos tal y como figura en la propia definición del caso y, aunque similar a los datos de una biblioteca, no está definido como tal.

Tabla 4. Subclases para Colecciones

Subclases	Número	Porcentaje
Bibliotecas	5	25%
Archivos	4	20%
Museos	2	10%
Bibliografías	4	20%
Colecciones	1	5%
Catálogos	4	20%
Total	20	100%

La siguiente clase definida en el documento elaborado por el LLD-XG hace referencia a los vocabularios de valores o vocabularios controlados. En esta ocasión a la hora de subdividir esta clase hemos empleado los mismos apartados que figuran recogidos en el propio documento. En este caso, aparecen con el mismo número tanto los sistemas de clasificación como los tesauros. Con dos ejemplos se presentan tanto los encabezamientos de materia como los datos de autoridad (Tabla 5).

Tabla 5. Subclases para Vocabulario de valores

Subclases	Número	Porcentaje
Sistemas declasificación	6	38%
Encabezamientos de materia	2	12%
Datos de autoridad	2	12%
Tesauros	6	38%
Total	16	100%

Finalmente, si subdividimos la última clase que se presenta en el VOCABDATASET denominada como conjunto de elementos de metadatos es significativo que la mayoría de los casos hacen referencias a ontologías. Tan sólo se da un único caso definido como metadatos que es nuevamente el Project Gutenberg pero en este caso presentado como RDF (Tabla 6).

Tabla 6. Subclases para Metadatos

Subclase	Número	Porcentaje
Ontologías	5	83%
Metadatos	1	17%
Total	42	100%

3.2. Descripción de los casos más representativos

En este apartado pasaremos a describir brevemente los casos más representativos de cada una de las clases anteriormente identificadas.

Comenzando por los *datasets* incluidos dentro de la primera clase establecida, la referente al apartado *Colecciones*, y dentro de ella la subclase *Bibliotecas*, destaca Europeana.

Europeana, actualmente contiene metadatos abiertos sobre 45 millones de textos, imágenes, vídeos y sonidos recogidos por esta institución. Estos objetos proceden de proveedores de datos que han reaccionado pronto y positivamente a la iniciativa de Europeana de promoción de datos más abiertos y nuevos acuerdos de intercambio de datos. Cubren una gran variedad de objetos del patrimonio en 45 idiomas: libros, periódicos, revistas, cartas, diarios, documentos de archivo, cuadros, pinturas, mapas, dibujos, fotografías, música, tradición oral grabada, emisiones de radio, películas y otros programas televisivos. Para el modelado de los datos siguen un esquema propio: *Europeana Data Model* (EDM). La adopción de este modelo basado en los estándares desarrollados por el W3C, ha permitido a Europeana la compatibilidad con el paradigma de la web semántica. EDM está constituido por la reutilización de *namespaces* como RDF (*Resource Description Framework)*, RDFs (*Resource Description Framework Schema*), OAI-ORE (*Open Archives Initiative Object Reuse and Exchange*), SKOS (*Simple Knowledge Organization System*) y DCMI Terms (*Dublin Core Metadata Iniciative Terms*) (Figura 2).

En el apartado denominado *Archivos* destaca por su importancia dentro de los datos vinculados la agrupación *Archives Hub Linked Data*. Este ejemplo recoge una muestra de datos de descripciones de fondos de archivos mantenidos en el *Archives Hub*, un agregador del Reino Unido, y emitidos como *linked data*. Ofrece una perspectiva sobre las personas, las organizaciones, las materias y lugares relacionados con los archivos que se describen. Los enlaces externos se proporcionan a otros conjuntos de datos, como el VIAF y la LCSH. Asimismo, proporciona una hoja de estilo para convertir los datos EAD (XML para archivos) en RDF XML.

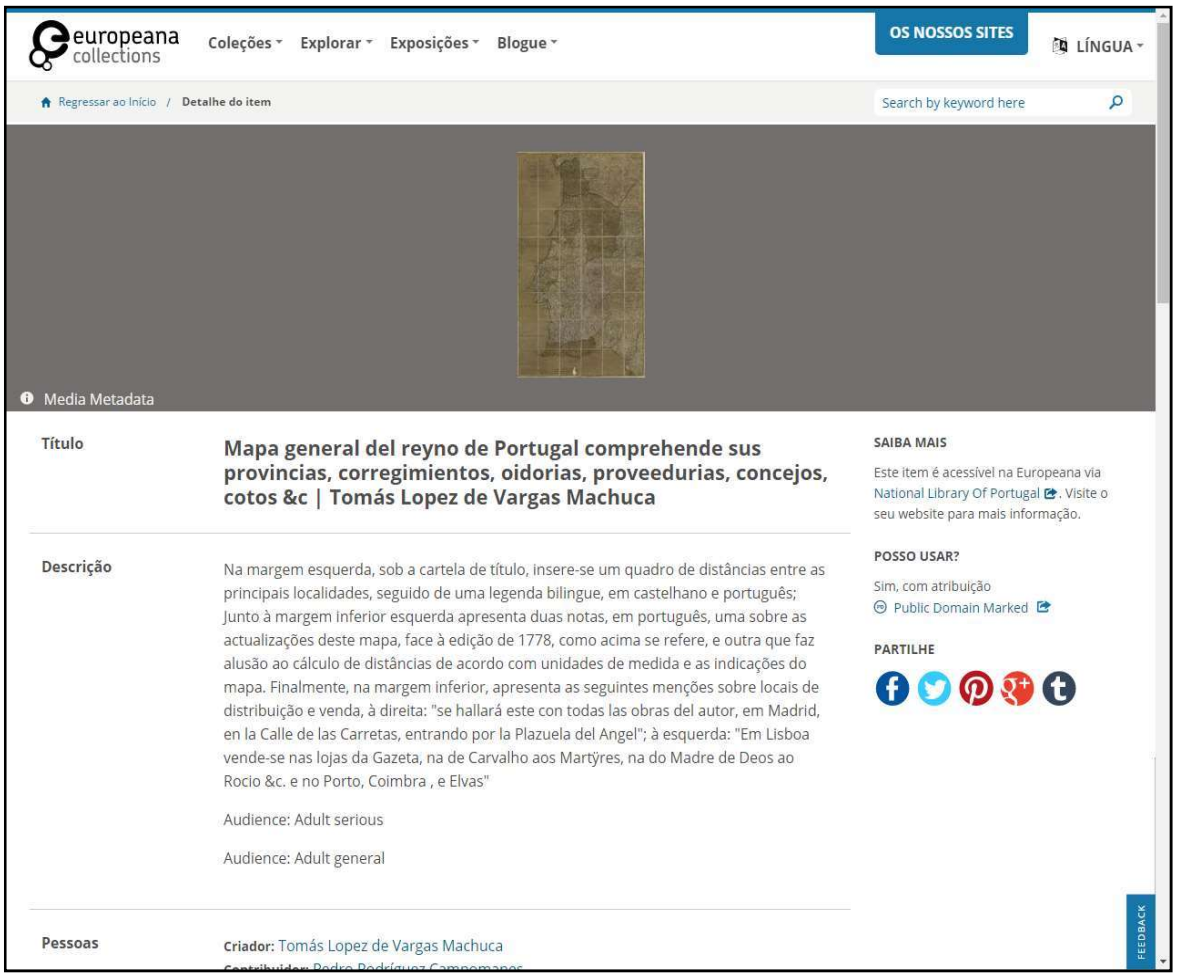

Figura 2. Ejemplo de Europeana para un mapa de Portugal

Dentro de la subclase catalogada como *Museos* es relevante la colección del *British Museum.* Este servicio de datos vinculados y SPARQL (*Protocol and RDF Query Language*) proporciona acceso a la misma colección de datos disponibles a través de la web del Museo. El uso del estándar de datos abiertos RDF del W3C, permite que los datos de la colección del museo se unan y se refieran a un creciente cuerpo de datos vinculados publicados por otras organizaciones de todo el mundo interesadas en la promoción de la accesibilidad y la colaboración. Los datos se han dispuesto mediante el empleo del CIDOC-CRM (modelo conceptual de referencia) fundamental para la armonización con otros datos del patrimonio cultural (Figura 3).

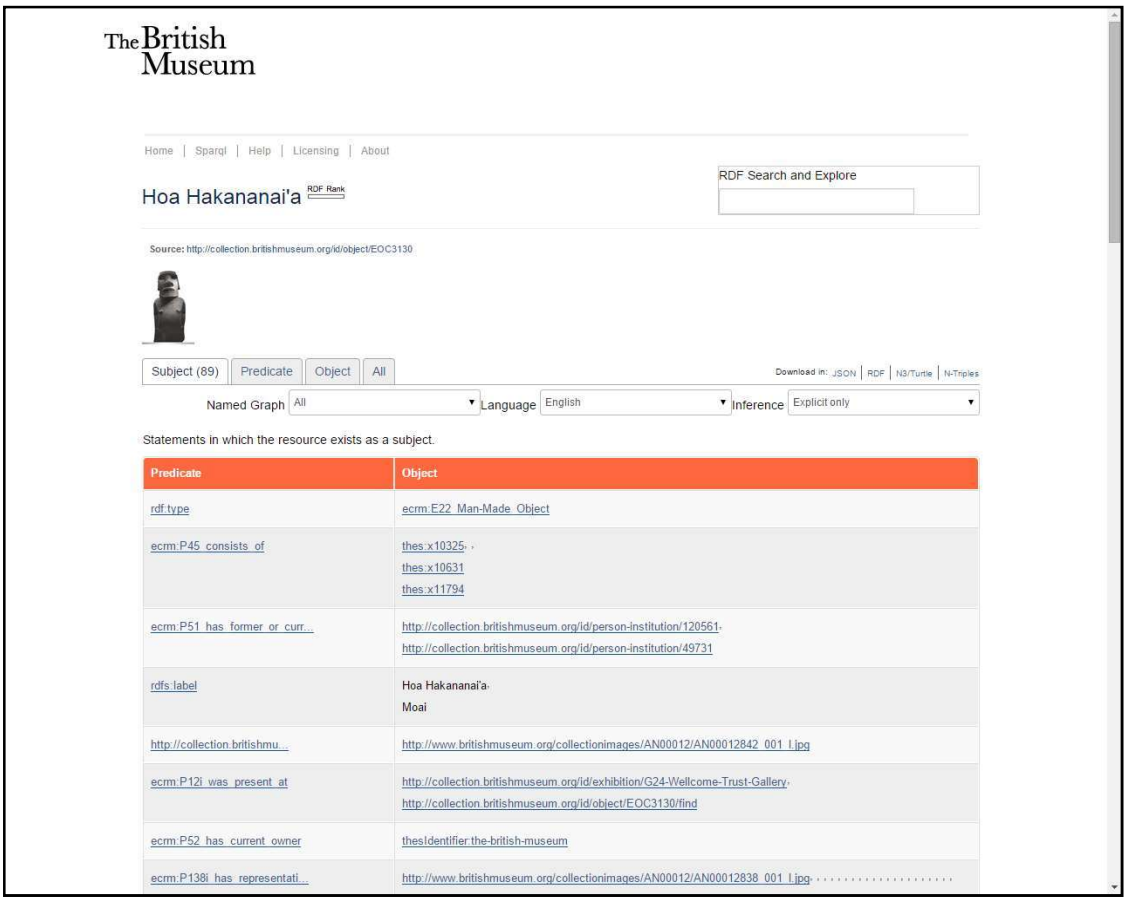

Figura 3. Ejemplo del Hoa Hakananai'a del British Museum en rdf

El *Proyecto Gutenberg*, cuya dificultad en su clasificación ya hemos descrito anteriormente, es la primera y más grande colección de libros electrónicos gratuitos. Sobre este caso no se proporciona información técnica en el catálogo *Mannheim* desde el punto de vista de ejemplo de aplicación LOD. Su versión en RDF figura como un caso específico dentro de la categoría metadatos, a la que posteriormente haremos referencia.

Si pasamos a la siguiente tipificación –*Vocabulario de valores*– comenzando por la subclase *Sistemas de clasificación* destaca la *Dewey Decimal Classification* (DDC). *Dewey.info* es un espacio experimental para los datos vinculados de la DDC. La intención de este prototipo es convertirse en una plataforma de los datos Dewey en la Web. Incluye como LD los sumarios DDC (los tres primeros niveles) de la vigésimo segunda edición, en once idiomas, y todos los números asignables de la decimocuarta edición abreviada, en tres idiomas. La clasificación semántica se codifica en RDF utilizando SKOS y otros vocabularios. Cada clase tiene también una representación HTML (XHTML + RDFa) y varios serializaciones RDF (RDF / XML, Turtle [*Terse RDF Triple Languag*], JSON [*JavaScript Object Notation*]).

Siguiendo con esta clase, en la subclase *Encabezamiento de materias* vamos a destacar por su interés, la *Library of Congress Subject Headings* (LCSH). La LCSH es la lista de materias más utilizada a nivel internacional. Desde el año 1898 la Biblioteca del Congreso ha mantenido de forma activa esta lista para la indización de sus fondos. Los modelos

utilizados para su implementación con LOD son: *Metadata Authority Description Schema* (MADS), esquema XML para los elementos de autoridad establecida por la propia institución, y SKOS (Figura 4).

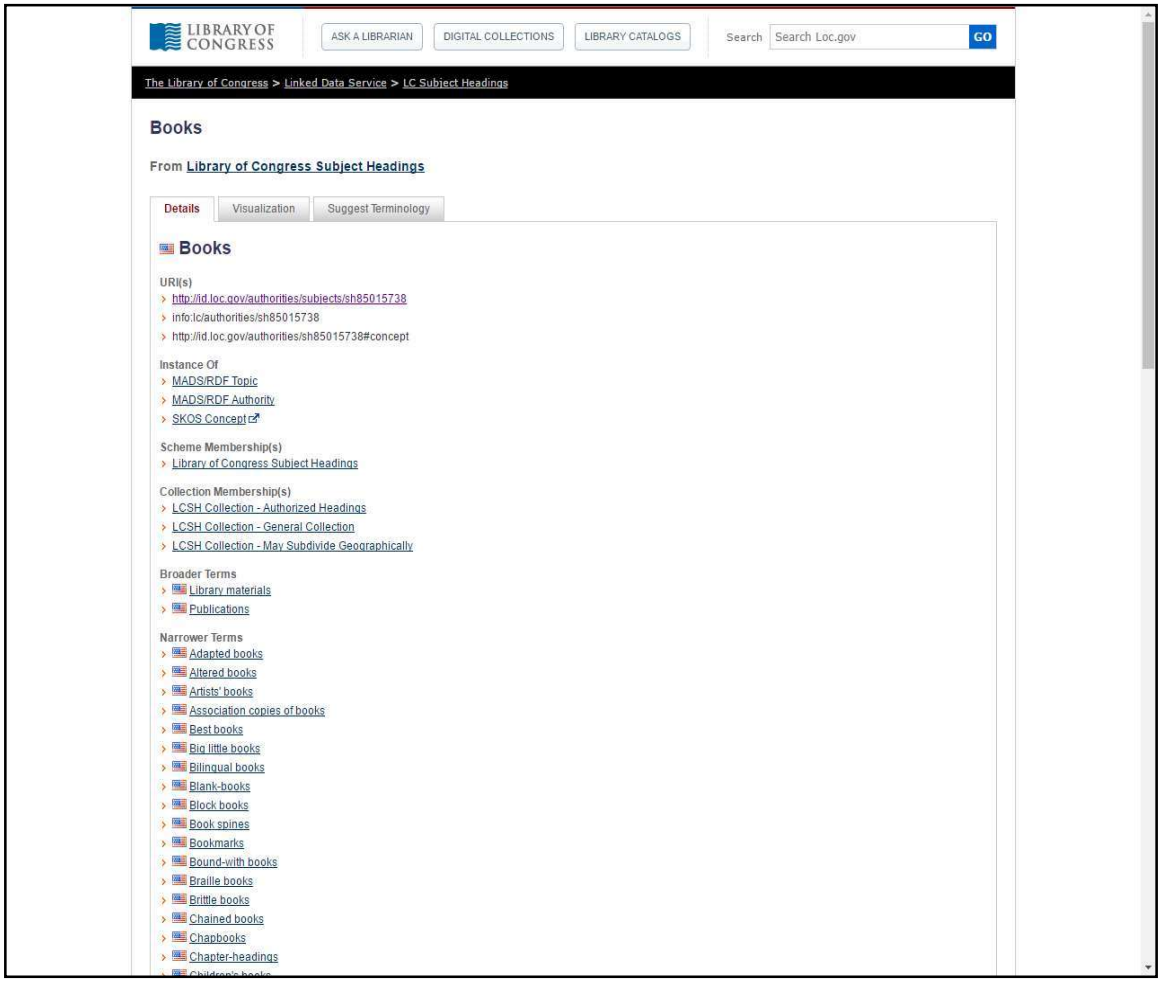

Figura 4. Ejemplo de la LCSH para la materia Libros

Finalmente, en el último apartado de esta categoría destacaremos el tesauro AGROVOC publicado por la FAO (Food and Agriculture Organization). *AGROVOC Linked Open Data (LOD)* es un proyecto para convertir el tesauro *AGROVOC* en una columna vertebral terminológica multilingüe para artículos digitales agrícolas. Alojado por MIMOS, socio en la investigación, provee acceso web a los registros de datos estructurados sobre concepto agrícola y, más importante aún, enlaces sobre esos conceptos a otros tesauros en línea. Consta de más de 32.000 conceptos disponibles en 24 idiomas. La versión de LD del tesauro está en RDF/ SKOS-XL, y el conjunto de tripletas se almacena en AllegroGraph (Figura 7).

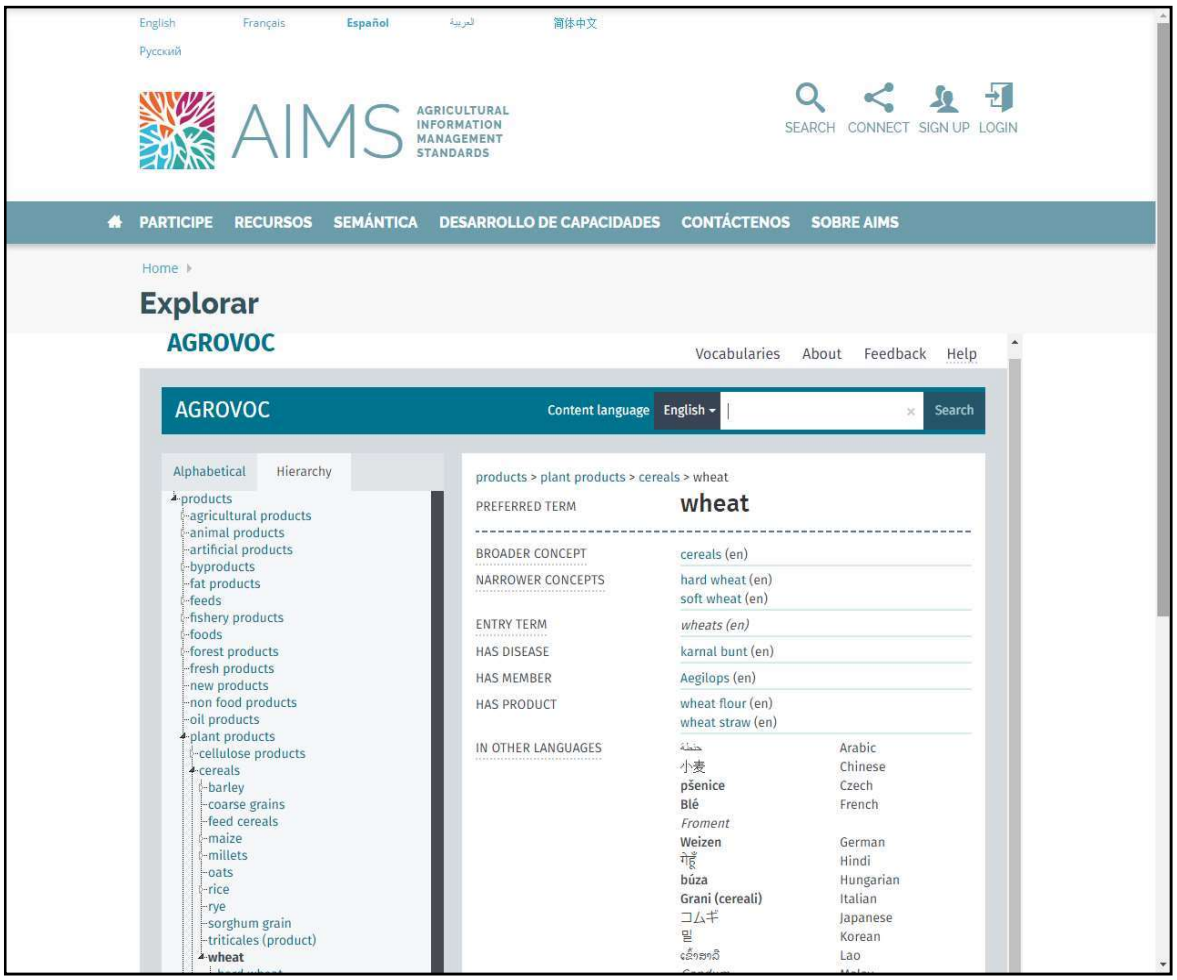

Figura 6. Ejemplo AGROVOC para el concepto Trigo

Para terminar con este apartado descriptivo concluiremos haciendo referencia a la última clase establecida –*Metadatos*– y al último caso nombrado: *Project Gutenberg in RDF*. Esta iniciativa ofrece los metadatos sobre las obras de dominio público disponibles en el *Proyecto Gutenberg* al que anteriormente ya habíamos hecho referencia dentro de la subclase *Colección*. El proyecto ha sido realizado por el grupo de investigación Research Group Data and Web Science de la Universidad de Mannheim. Los datos estructurados se proporcionan de acuerdo a los principios de LD y se pueden acceder a los mismos a través del sitio web del grupo por medio de tres vías diferentes: HTML, RDF y SPARQL (Figura 8).

4. Conclusiones

En un estudio previo (Ríos; Gil 2015) subrayábamos las dificultades a las que habíamos tenido que hacer frente a la hora de realizar nuestro estudio y las anomalías encontradas, tanto en la configuración del *LOD cloud diagram* como en la estructura del catálogo *Mannheim*, con el objetivo de que pudieran ser solventadas en las próximas ediciones de ambos instrumentos repercutiendo en una susceptible mejora de los mismos.

Con la nueva actualización de la "nube" tenemos que decir que no sólo no se han solucionado los problemas anteriormente referidos sino que éstos se han agudizado. Tal es así que, en primer lugar, no se ha realizado un análisis estadístico pormenorizado de cada uno de los *datasets*. Por otro lado, se ha dejado de actualizar el catálogo *Mannheim,* y la única fuente disponible para acceder a la información de cada conjunto es el Data Hub. Sin embargo, al no haber una categorización previa por categoría de *datasets,* por un lado, y al haber aumentado de modo considerable el número de conjuntos, por otro, hace que volver hacer un nuevo recuento se convierta en una tarea ardua y compleja. Con lo cual los datos aquí presentados son aproximados al haber tomado como fuente la anterior versión del *cloud diagram,* No obstante, y como ya hemos comentado a lo largo de estas páginas, nuestro objetivo no era tanto la cuantificación sino la clasificación de los subconjuntos de datos agrupados bajo la denominación de *Publicaciones*.

A este respecto, pensamos que la denominación de la categoría *Publicaciones* no es muy acertada. Si bien es cierto que los autores de la nube han querido englobar en un número reducido las categorías que comprende el diagrama y que esta denominación es un concepto muy amplio y en ella se pretende reunir la información científica y cultural. Nuestra propuesta sería o bien cambiar la denominación con un epígrafe similar al anteriormente indicado o utilizar dentro de esta categoría etiquetas más específicas que nos permitan encontrar directamente las distintas subcategorías. En este sentido para los conjuntos de datos GLAM, proponemos la misma clasificación que hemos empleado nosotros cuya definición viene avalada por el W3C.

Por otro lado, la información individualizada de cada *dataset* que nos proporciona el catálogo *Mannheim* es muy dispar y en ocasiones muy escueta. De tal modo que en un número importante de casos hemos tenido que proceder a buscar el enlace del mismo teniendo en cuenta que no es igual la dirección de una biblioteca, por ejemplo, que la dirección al proyecto de datos enlazados de esa misma institución. Es decir, la URL de Europeana es *www.europeana.eu* mientras que su proyecto de datos enlazados es *data.europeana.eu*. Incluso hay casos en los que ni siquiera se proporciona información por lo que hemos tenido que llevar a cabo toda una tarea de investigación para identificar el ejemplo en concreto y así poderlo categorizar. El responsable último de este hecho no es tanto el propio catálogo sino los responsables de los *datasets*. Por lo tanto, abogamos porque se proporcione la mayor información posible por parte de los gestores de los conjuntos de datos en aras a facilitar tanto el uso del propio *dataset* como el análisis del mismo.

Uno de los objetivos que nos planteábamos en el inicio del trabajo era la definición de los conceptos claves tanto en lo concerniente a la tecnología de los datos abiertos y vinculados como a la definición de conceptos más específicos de nuestro ámbito, en concreto, los referentes a los recursos bibliotecarios empleados: colecciones, vocabularios de valores y metadatos. A través de la consulta de la documentación especializada sobre el tema comprobamos que en ocasiones ambos conceptos se confunden y se está perdiendo precisión terminológica. Mediante la explicación de los términos en el apartado teórico y la sistematización de los ejemplos que figura en el apéndice esperamos haber cumplido con nuestro propósito.

Tras el análisis estadístico, cuyos resultados ya hemos comentado en el apartado previo, en la última parte hemos intentando describir los casos más representativos de las categorías previamente establecidas, centrándonos no sólo en

su definición si no en la aplicación de las tecnologías de datos vinculados. Es así que las instituciones culturales que estén interesadas en publicar sus dados como LOD obtengan con este informe una primera aproximación a la aplicación de esta tecnología, remitiendo a cada caso particular para obtener una información más exhaustiva.

Sobre este punto diremos que nos hubiera gustado profundizar en la explicación de los recursos empleados a la hora de disponer los datos como LOD. Sin embargo este hecho no ha sido posible debido a la dificultad anteriormente comentada para encontrar la información referente a la utilización de las diferentes prácticas relativas a este aspecto.

5. Fuentes de información

BAUER, F.; KALTENBÖCK, M. Linked Open Data: The Essentials: a Quick Start Guide for Decision Makers. Viena, Austria: edition mono/monochrom, 2012.

BERNES-LEE, Tim. Linked data – Design Issues [em linha]. [S.l.]: Bernes-Lee, 2006. [Consult. 1 jun. 2017]. Disponível na internet:http://www.w3.org/DesignIssues/LinkedData.html.

EUROPEANA Linked Open Data [em linha]. European Union: Europeana Fundation: Europenan Creative Project, 2015. [Consult. 1 jun. 2017]. Disponível na internet: http://data.europeana.eu.

GUÍA Breve de Linked Data [em linha]. Gijón: W3C Oficina España, 2015. [Consult. 1 jun. 2017]. Disponível na internet: http://www.w3c.es/Divulgacion/GuiasBreves/LinkedData.

GUÍA breve de Web Semántica [Em linha]. Gijón: W3C Oficina España, 2017. [Consult. 1 jun. 2017]. Disponível na internet: http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica.

HEATH, T.; BIZER, C. Linked data: Evolving the Web into a Global Data Space. Florida, USA: Mor-gan & Claypool Publishers, 2011.

ISAAC, Antoine; WAITES, William; YOUNG, Jeff; ZENG, Marcia. Library Linked data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets [em linha] [S.l]: W3C, 2011. [Consult. 1 jun. 2017]. Disponível na internet: http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025.

LIBRARY Linked Data Group Incubator. Linked Data Incubator Group Final Report [em linha] [S.l.]: World Wide Web Consortium, 2011. [Consult. 1 jun. 2017]. Disponível na internet: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/.

LIBRARY Linked Data Group Incubator. Linked Data Incubator Group Use Case [em linha]. [S.l.]: World Wide Web Consortium, 2011. [Consult. 1 jun. 2017]. Disponível na internet: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/#ref-USECASE.

LIBRARY terminology informally explained [Em linha] . [S.l.]: World Wide Web Consortium, 2011. [Consult. 1 jun. 2017]. Disponível na internet: http://www.w3.org/2001/sw/wiki/Library_terminology_informally_explained#Definitions.

LINKED data: connect Distributed Data across the Web [En linha]. [S.l.]: [s.n.], 2014. [Consult. 1 jun. 2017]. Disponível na internet: http://linkeddata.org.

LINKING Open Data [En linha]. [S.l.]: World Wide Web Con-sortium, 2017. [Consult. 1 jun. 2017]. Disponível na internet: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

MANNHEIM Linked Data Catalog [em linha]. Mannheim: University of Mannheim, 2014. [Consult. 1 jun. 2017]. Disponível na internet: http://linkeddatacatalog.dws.informatik.uni-mannheim.de (2015-03-30).

MITCHEL, Erik -- Library Linked data: research and adoption. Library Technology Reports. Vol. 5, nº 49 (2013), p. 11-25.

OPEN Definition Advisory Council [em linha]. Cambridge: Open knowledge, 2014. [Consult. 1 jun. 2017]. Disponível na internet: http://opendefinition.org/.

SCHMACHTENBERG, Max; BIZER, Christian; PAULHEIM, Heiko Adoption of the linked data best practices in different topical domains [em linha]. Mannheim: University of Mannheim; Unión Europea: Planet Data, 2014a. [Consult. 1 jun. 2017]. Disponível na internet: http://dws.informatik.uni-mann-heim.de/ fileadmin/lehrstuehle/ki/pub/SchmachtenbergBizerPaulheim-AdoptionOfLinkedDataBestPractices.pdf.

RÍOS Hilario, Ana B.; GIL Urdiaciain, Blanca - Los datos bibliotecarios en la nube de datos: Análisis de los datasets GLAM presentes en el LOD cloud diagram. Scire. Vol. 17, nº 2 (jul.-dic. 2011) p. 35-47.

SCHMACHTENBERG, Max; BIZER, Christian; PAULHEIM, Heiko [em linha]. State of the LOD Cloud 2014. Mann-heim: University of Mannheim; Unión Europea: Planet Data, 2014b. [Consult. 1 jun. 2017]. Disponível na internet: http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/#toc1.

Apéndice

Listado de los conjuntos de datos bibliotecarios del *LOD cloud diagram*

Colecciones	**Bibliotecas**	
	Princeton Library FindingAids	http://findingaids.princeton.edu/
	data.bnf.fr- Bibliothèque nationale de France	http://data.bnf.fr/
	datos.bne.es	http://datos.bne.es
	Europeana Linked Open Data	http://data.europeana.eu/
	Public Library of Veroia	http://libver.math.auth.gr
	Archivos	
	Airforcehistoryindex	http://airforcehistoryindex.org/
	Muninn World War I	http://rdf.muninn-project.org
	Archives Hub Linked Data	http://data.archiveshub.ac.uk/
	National Digital Data Archive of Hungary	http://lod.sztaki.hu/
	Museos	
	British Museum Collection	http://collection.britishmuseum.org/
	Amsterdam Museum	http://semanticweb.cs.vu.nl/lod/am/
	Bibliografías Nacionales	
	British National Bibliography	http://bnb.data.bl.uk/
	Deutsche Biographie	http://www.deutsche-biographie.de/
	Deutsche Nationalbibliografie	http://www.dnb.de/EN/datendienste/linkedData
	LIBRIS	http://libris.kb.se
	Colecciones	
	Project Gutenberg	http://www.gutenberg.org/
	Catálogos	
	Sudoc bibliographic data	http://punktokomo.abes.fr/2011/07/04/le-sudoc-sur-le-web-de-donnees/
	Hungarian National Library (NSZL) catalog	http://nektar.oszk.hu/wiki/Semantic_web

Vocabularios de valores	HeBIS	http://www.hebis.de/
	85ZDB	http://www.zeitschriftendatenbank.de/services/schnittstellen/linked-data/
	Sistemas de clasificación	
	Mathematics Subject Classification	http://msc2010.org/mscwork/
	Faceted Application of Subject Terminology	http://experimental.worldcat.org/fast/
	Dewey Decimal Classification (DDC)	http://dewey.info
	JITA Classification System of Library and Information Science	http://www.destin-informatique.com/ASKOSI/Wiki.jsp?page=JITA%20Maintenance
	ICONCLASS- Multilingual Thematic Classification	http://www.iconclass.org/help/lod
	Glottolog	http://glottolog.org
	Encabezamientos de materia	
	Lista de Encabezamientos de Materia	http://id.sgcb.mcu.es
	Library of Congress Subject Headings	http://id.loc.gov/authorities/
	Datos de autoridad	
	IdRef: Sudoc authority data	http://punktokomo.abes.fr/2011/07/05/idref-des-pages-html-et-rdf-plus-riches/
	Gemeinsame Normdatei (GND)	http://d-nb.info/standards/elementset/gnd#
	Tesauros	
	yso-fi-allars	http://finto.fi/allars/en/page/Y46304?rdf=xml
	Unesconom	http://skos.um.es/unesco6/6101/rdfxml
	STW Thesaurus for Economics	http://zbw.eu/stw/versions/latest/about
	National Agricultural Library Thesaurus	http://agclass.nal.usda.gov/agt.shtml
	AGROVOC	http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus
	Mis Museos, índice semántico de museos, artistas y obras de arte (GNOSS)	http://mismuseos.net/comunidad/metamuseo
	Ontologías	
	yso-fi-ysa	http://www.yso.fi/onto/ysa/Y99974?rdf=xml
	Biblie Ontology	http://bibleontology.com/
	ASN:US	http://asn.jesandco.org
Metadatos	Traditional Korean Medicine Ontology	http://tkm.kiom.re.kr
	FAO geopolitical ontology	http://www.fao.org/countryprofiles/geoinfo/en/?lang=en
	Metadatos	
	Project Gutenberg in RDF	http://www4.wiwiss.fu-berlin.de/gutendata/