ENRIQUECIMIENTO SEMÁNTICO DE COLECCIONES DIGITALES A TRAVÉS DEL EUROPEANA DATA MODEL

César Juanes Hernández
Dpto. Innovación de DIGIBÍS
C/ Alenza, 4, planta 5. 28003 Madrid – España
Teléfono: (+34) 914 320 888, Fax: (+34) 914 321 113
cesar.juanes@digibis.com

RESUMEN

El modelo de datos de Europeana (EDM) se ha convertido en el resorte ideal para la aplicación de tecnología Linked Open Data en bibliotecas, archivos y museos. Esta comunicación describe, a partir del *White Paper* publicado por Europeana el 30 de mayo de 2015 y titulado *Enhancing the Europeana Data Model (EDM)*, la evolución del modelo de datos durante los últimos cinco años y los conceptos básicos para su implementación en las colecciones digitales de bibliotecas, archivos y museos.

Para analizar la evolución del modelo, se enumeran diferentes colecciones digitales españolas como ejemplos de buenas prácticas de aplicación y uso, y se describen diferentes procedimientos de enriquecimiento semántico a través del uso de conjuntos de datos y vocabularios Linked Open Data.

PALABRAS CLAVE

Europeana Data Model; Bibliotecas Digitales; Linked Open Data; Europeana; Enriquecimiento semántico.

INTRODUCCIÓN

Bajo el título *The Europeana Data Model: A living model 5 years on* (Isaac, 2015), Europeana daba a conocer el 17 de junio de 2015, a través de su Blog, la publicación de su, por entonces, nuevo *White Paper*. El objetivo, destacar la evolución del Europeana Data Model (EDM) e incidir en los principales puntos de acción que han convertido a este modelo de datos en el requisito imprescindible que los recursos digitales deben cumplir para su agregación en Europeana.

El *White Paper*, titulado *Enhancing the Europeana Data Model (EDM)* (Charles e Isaac, 2015) refleja la evolución experimentada por el propio modelo de datos en los

últimos cinco años, desde la publicación de la versión 5.1 a principio de 2010, hasta la actual versión 5.2.6 del modelo, e incide en la flexibilidad como una de sus características principales puesto que, al no circunscribirse a una tipología documental concreta, puede acomodarse perfectamente al ámbito bibliotecario (con descripciones en formato MARC), archivístico (con descripciones en EAD) o museístico (con descripciones en formato LIDO).

EDM se ha convertido en cinco años en la pieza fundamental dentro de la operativa de Europeana que agrega, procesa y enriquece los recursos digitales de bibliotecas, archivos y museos de toda Europa. Su desarrollo y consolidación han relegado a un segundo lugar al anterior esquema requerido por Europeana, el denominado Europeana Semantic Elements (ESE), definido como un modelo de descripción plano, que utilizaba cadenas de texto literales en la gran mayoría de sus elementos y que impedía la vinculación de recursos de una forma efectiva, para dar paso a un modelo orientado hacia la tecnología Linked Open Data. Gracias al uso de esta tecnología se hace posible interoperar entre recursos de diferentes instituciones y relacionar los datos de las colecciones con otros datos disponibles en la Web basándose en los principios generales de la Web Semántica a través el uso de URIs HTTP para identificar recursos, la utilización de RDF o la inclusión de enlaces a otros URIs.

Lejos de ser considerado como un requisito más dentro de las prescripciones técnicas que deben ser cumplidas por los proveedores de datos de la plataforma Europeana, EDM se erige como una oportunidad, un primer paso, para la semantización de bases de datos de bibliotecas, archivos y museos. A través de su demostrada capacidad para establecer estructuras de navegación estables, tanto en grandes colecciones digitales, como en colecciones locales de menor tamaño, se marca el objetivo de proporcionar información contextual y de esta forma poder establecer relaciones y equivalencias entre diferentes recursos de distinta procedencia, ofreciendo así recursos informativos de mayor calidad.

UN MODELO DE DATOS FLEXIBLE Y SEMÁNTICO PARA BIBLIOTECAS, ARCHIVOS Y MUSEOS

EDM, como modelo de datos, define las clases y propiedades que deben ser utilizadas al estructurar una colección digital de una biblioteca, un archivo o un museo. Ha sido diseñado para una explotación semántica de los datos y su objetivo es permitir la integración de los diferentes formatos de descripción en el ámbito del patrimonio cultural de forma que las descripciones de origen puedan ser recopiladas y conectadas a través de conceptos. EDM no es un modelo de descripción y por lo tanto la codificación de los registros puede y debe realizarse a través de los formatos propios de codificación de cada ámbito como son, por ejemplo, el formato MARC para bibliotecas, EAD para archivos, o LIDO para museos.

La fortaleza de este modelo radica en haber sido diseñado, no como un modelo de datos propio para Europeana, sino como un modelo de datos para bibliotecas, archivos y museos que desean participar en Europeana y establecer un carácter semántico a sus recursos informativos. Teniendo en cuenta las diferencias existentes en las políticas de gestión de colecciones de bibliotecas, archivos y museos, resulta esencial la flexibilidad de EDM para encontrar acomodo en colecciones tan heterogéneas y así obtener un alto grado de aceptación en la implementación del modelo. El desarrollo de numerosos proyectos a través de la aplicación de EDM como eje vertebrador de colecciones digitales, como la *Deutsche Digitale Bibliothek*[1], el proyecto *Digitised Manuscripts to Europeana (DM2E)*[2], *Musical Instrument Museums Online (MIMO)*[3] o, muy especialmente, la *Digital Public Library of America (DPLA)*[4], cuyo modelo de datos, *Metadata Application Profile v.4.0* (2015) ha sido elaborado a partir de la experiencia del modelo de datos de Europeana, ponen de manifiesto hasta qué punto EDM puede encontrar encaje en cualquier tipo de colección digital.

Además de la flexibilidad, la característica fundamental de EDM es su marcado carácter semántico y su orientación hacia la aplicación de políticas y tecnología Linked Open Data. En este sentido, Europeana, con más de 44 millones de recursos digitales y DPLA, con más de 11 millones, son un claro referente de colecciones que reúnen recursos digitales de diferente tipología, procedencia o ámbito lingüístico. El crecimiento de estas dos plataformas pone de relieve la necesidad de disponer de un modelo de datos adecuado a la tecnología Linked Open Data, que permita vincular recursos y ofrecer información contextual para mejorar las capacidades semánticas de la información accesible en la Web. Esta necesidad queda perfectamente representada en los documentos *Europeana Strategy 2015 – 2020* (Cousins, Poole y Racine, 2013), donde se enumeran las prioridades de Europeana para ese intervalo de años y en el que, en primer lugar, aparece la necesidad de seguir invirtiendo en estructuras Linked Open Data; o el *Strategic Plan de la DPLA* (2015) en el que también se establece como prioridad mejorar los metadatos a través del desarrollo de políticas Linked Data.

EDM: APLICACIÓN Y USO

EDM ha experimentado un importante desarrollo en los últimos cinco años. Toda la información sobre el modelo de datos y la evolución experimentada se encuentra accesible a través del apartado Web *Europeana Professional*[5] desde el que se tiene acceso a la última versión del modelo de datos (la 5.2.6) en la que se describen las particularidades del mismo.

[1] Accesible a través de https://www.deutsche-digitale-bibliothek.de/. Consulta: 28/09/2015
[2] Accesible a través de http://dm2e.eu/. Consulta: 28/09/2015
[3] Accesible a través de http://www.mimo-international.com/MIMO/. Consulta: 28/09/2015
[4] Accesible a través de http://dp.la/. Consulta: 28/09/2015
[5] Accesible a través de http://pro.europeana.eu/. Consulta: 28/09/2015

A nivel nacional, es importante destacar la elaboración del *Manual Básico de Europeana Data Model* que, aunque no es una versión oficial del *Europeana Data Model Primer* (Claypham e Isaac, 2013), contiene la explicación de cómo deben utilizarse conjuntamente las clases y propiedades para modelar los datos y sustentar la operativa de Europeana.

Pese a los cambios producidos, la esencia de EDM se mantiene intacta, puesto que su intención ha sido siempre la de enriquecer las descripciones de los objetos y contextualizarlos a través de tecnología Linked Open Data.

Con este objetivo, EDM define cinco clases dedicadas a la representación de entidades contextuales (Claypham, Charles e Isaac, 2014) a través de las cuales establece un marco de actuación para la publicación de datos que ofrezcan un giro semántico.

- Agent: entidad utilizada para representar personas y organizaciones.
- Event: entidad utilizada para representar acontecimientos.
- Place: entidad utilizada para entidades espaciales.
- TimeSpan: entidad utilizada para representar periodos de tiempo y fechas.
- Concept: entidad utilizada para representar sistemas de organización como tesauros, esquemas de clasificación o encabezamientos de materias.

La descripción de estas entidades contextuales está enfocada a centrar la navegación en las personas, los eventos, los lugares, las fechas y las materias con el objetivo de responder a las preguntas ¿Quién?, ¿Qué?, ¿Cómo?, ¿Cuándo? y ¿Dónde?, tal y como queda reflejado en el propio formulario de búsqueda de Europeana y en su navegación por facetas.

Identificadas las entidades contextuales sobre las que se debe centrar el proceso de semantización de la base de datos, bibliotecas, archivos y museos deben integrar en su política de descripción el uso de conjuntos de datos y vocabularios Linked Open Data. El número de conjuntos de datos y vocabularios de valores se ha incrementado progresivamente en los últimos años, tal y como recogen Agenjo y Hernández (2015) en su análisis sobre el estudio que la OCLC publicó en 2014 sobre la implementación de Linked Data. A partir de ese análisis, esta comunicación ofrece diez ejemplos de conjuntos de datos y vocabularios que, por sus características, pueden ser empleados indistintamente por colecciones digitales de bibliotecas, archivos y museos.

- **id.loc.gov: http://id.loc.gov/**

Creado en el año 2009, tiene su origen en el servicio Linked Data de la *Library of Congress*. Ofrece diferentes modalidades de descarga de ficheros, interfaz REST o búsqueda Open Search.

- **DBpedia. http://id.loc.gov/**

Proyecto iniciado en 2007 para extraer de la *Wikipedia* la estructura semántica de las distintas entradas textuales que contiene. Esta fuente puede ser utilizada para enriquecer descripciones de personas, lugares u organizaciones.

- **GeoNames. http://www.geonames.org/**

Utilizada para ofrecer información contextual de lugares, incorpora información semántica de 8.3 millones de topónimos.

- **Virtual International Authority File (VIAF). http://viaf.org/**

Base de autoridades, principalmente de tipo persona o institución, generada por la OCLC a partir de los ficheros de autoridades de múltiples bibliotecas nacionales y organizaciones. Los datos están disponibles con una licencia ODC attribution para ser descargados en diversos formatos como RDF-XML, RDF-NT, ISO2709, MARC-XML.

- **Faceted Application of Subject Terminology (FAST). http://fast.oclc.org/**

Contiene enlaces a *Library of Congress Subject Headings* (LCSH) y a otras fuentes autorizadas como VIAF, GeoNames y Wikipedia, así como a WorldCat, con indicación del número de veces que se utiliza un determinado término.

- **AAT, ULAN y TGN. http://vocab.getty.edu/**

Estos tres conjuntos de vocabularios, *Art & Architecture Thesaurus* (AAT), *Union List of Artist Names* (ULAN) y *Thesaurus of Geographic Names* (TGN) han sido publicados en Linked Open Data por la *Paul Getty Foundation* bajo licencias Open Data Commons Attribution License (ODC-By) v1.0 y cuentan con la ventaja de disponer de un servidor SPARQL a través de cual descargar la información.

- **Datos.bne.es. http://datos.bne.es/**

Portal de datos de la *Biblioteca Nacional de España* que cuenta con información accesible a través de la licencia CC0 (Creative Commons Public Domain Dedication).

- **Lista de Encabezamientos de Materia para Bibliotecas Públicas (LEM): http://id.sgcb.mcu.es/lem/**

Proyecto de la *Subdirección General de Coordinación Bibliotecaria* que ha cruzado sus términos con la *Lista de encabezamentos de materia en galego* (LEMAG) y la *Llista de encapçalaments de matèria de la Biblioteca de Catalunya* (LEMAC), además de establecer vínculos a otras listas de encabezamientos de materia como la *Library of Congress Subject Headings* (LCSH), RAMEAU o *Gemeinsame Normdatei* (GND). Sus datos pueden ser descargados en RDF/XML o en formato MARC 21 para la importación directa en cualquier sistema de gestión. Además, cuenta con un servidor SPARQL para la interrogación y descarga de sus datos.

VINCULAR, ENRIQUECER Y CONTEXTUALIZAR.

Para la implementación de este modelo de datos, bibliotecas, archivos y museos se enfrentan a la necesidad de emprender una tarea de enriquecimiento semántico en sus colecciones digitales haciendo uso de conjuntos de datos y vocabularios Linked Open Data.

Si bien el uso de un modelo de datos y su adecuación a los principios Linked Open Data puede englobarse dentro de las capacidades tecnológicas de una base de datos o, más bien, de las capacidades del software de gestión, el enriquecimiento semántico propiamente estaría dentro de las características funcionales y de procedimiento del proceso de descripción.

La tarea de enriquecer semánticamente una colección digital puede definirse como el proceso que, mediante la realización de búsquedas en diferentes conjuntos de datos y vocabularios, o contra conjuntos de datos ya descargados en el sistema, permite enriquecer la descripción de los objetos gestionados incluyendo en el campo, subcampo o etiqueta correspondiente, la URI que identifica de forma inequívoca los conceptos.

Lógicamente, estos procesos de enriquecimiento pueden realizarse de forma desatendida (o automática) para enriquecer un volumen grande de registros, por ejemplo utilizando APIs que permiten extraer los conceptos del vocabulario analizado y vincularlo a los conceptos almacenados en nuestra base de datos, o de forma interactiva (manual) a petición del usuario desde la página de edición de un objeto.

Cualquiera que sea el procedimiento utilizado, el fin último consiste en incluir en cada una de las entidades enriquecidas una URI que permita identificar, contextualizar y, en última instancia, vincular con otras entidades relacionadas.

Para llevar a cabo el enriquecimiento semántico de las descripciones bibliográficas, archivísticas o museísticas, el formato empleado para la descripción debe incorporar un elemento, campo o etiqueta a través del cual se pueda incluir el vínculo a recursos Linked Open Data.

El ejemplo más representativo de enriquecimiento semántico es el realizado sobre registros codificados según el formato MARC, que disponen de un campo, el 024 (Otros identificadores normalizados), en el que es posible registrar la URI del concepto. De esta forma, con el simple procedimiento de incluir en el campo 024 de los registros de autoridad la URI de un recurso Linked Open Data, se procede al enriquecimiento de la entidad contextual.

Este es el procedimiento empleado, por ejemplo, por la *Biblioteca Virtual de la Provincia de Málaga*[6] que, a través de su sistema de gestión, procede a la integración sistemática de URIs en los registros de autoridad de la base de datos codificados según el formato MARC.

[6] Accesible a través de http://bibliotecavirtual.malaga.es/. Consulta: 28/09/2015

Así, la colección de *Personajes malagueños*, ofrece un listado de autores descritos en formato MARC que incorporan un campo 024 haciendo referencia a vocabularios como los del VIAF o FAST.

Figura 1. Ejemplo de registro de autoridad (Persona) con enriquecimiento semántico.

EDM EN ESPAÑA: EJEMPLOS DE BUENAS PRÁCTICAS

A nivel nacional, la aportación de recursos digitales a Europeana se puede calificar de excelente, y así lo refrendan los datos estadísticos que posicionan a España como el cuarto proveedor de datos totales a Europeana y, de forma más concreta, a *Hispana*[7], *Directorio y Recolector de colecciones digitales del Ministerio de Educación, Cultura y Deporte*, como el cuarto proveedor de datos[8] totales a la plataforma europea.

Esta posición privilegiada se ve fortalecida por la creación de nuevas colecciones digitales. Una de las últimas, la *Biblioteca Digital AECID*[9], presentada oficialmente el pasado 21 de mayo de 2015 con motivo del Acto Conmemorativo de los 75 años de la Biblioteca de la AECID y que ofrece la posibilidad de consultar, a través del modelo de datos de Europeana, más de 2.000 obras y 1.000.000 de imágenes de tres grandes colecciones como la *Biblioteca Hispánica*, la *Biblioteca Islámica* y las *Publicaciones de la AECID*, que fueron recolectadas por Europeana tan solo unas pocas semanas después de su presentación oficial.

[7] Accesible a través de http://hispana.mcu.es/. Consulta: 28/09/2015
[8] Accesible a través de http://www.europeana.eu/portal/europeana-providers.html. Consulta: 28/09/215
[9] Accesible a través de http://bibliotecadigital.aecid.es/. Consulta: 28/09/2015

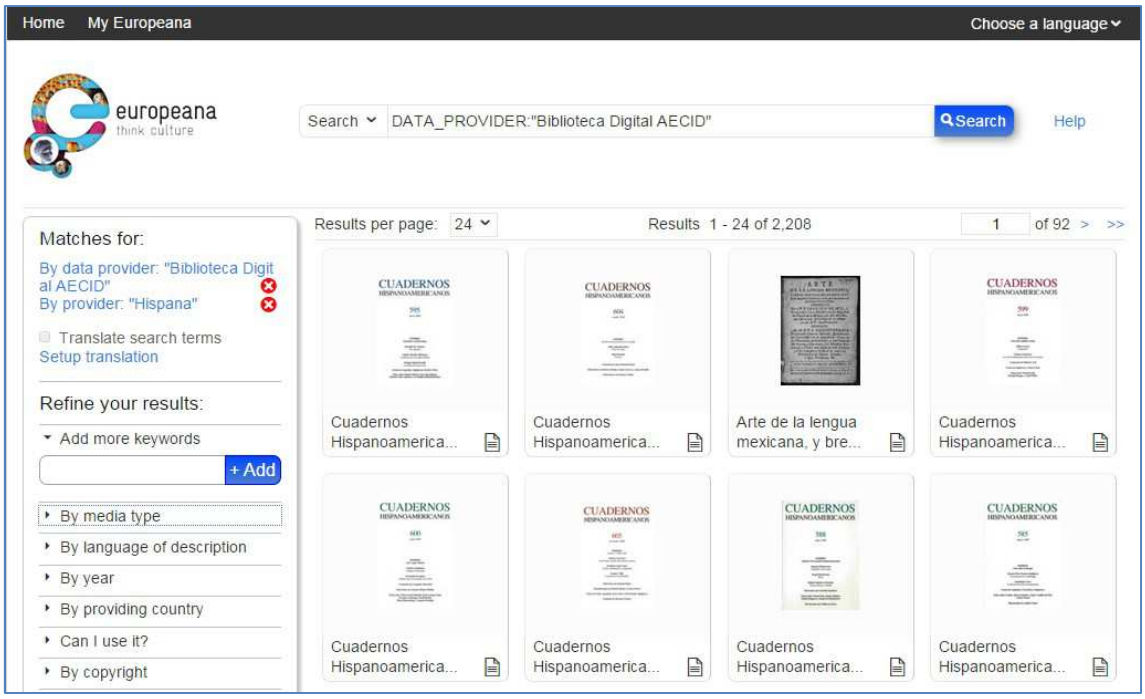

Figura 2. Aportación de la Biblioteca Digital AECID a Europeana.

No obstante, la aportación de las colecciones digitales españolas no está igual de valorada dentro de Europeana, obteniendo más protagonismo aquellas que implementan de forma efectiva EDM, en lugar de las que siguen implementando el antiguo esquema ESE.

Así, en el apartado Web *Europeana Professional*, la Web común de los proyectos de Europeana, se destacan diferentes casos de estudio de colecciones digitales. Para EDM, la *Biblioteca Virtual Ignacio Larramendi de Polígrafos*[10] (The Polymath Virtual Library, en inglés) ocupa un lugar privilegiado. Este proyecto, desarrollado por la *Fundación Ignacio Larramendi* da buena cuenta del papel protagonista de España en la construcción de Europeana, siendo designado como caso de estudio de EDM (Europeana Foundation, 2012).

Otro ejemplo reconocido como modelo de buenas prácticas en la implementación de EDM lo representa la *Biblioteca Virtual del Ministerio de Defensa*. Así queda recogido en el *Europeana Labs*[11], plataforma en la que se seleccionan algunas de las colecciones más destacadas que forman parte de Europeana atendiendo al uso de licencias abiertas que permiten la reutilización de contenidos y también a la calidad de los datos proporcionados, donde se ha hecho público un *data set* perteneciente a la colección de mapas y planos del Ministerio de Defensa que incide en la excelente práctica catalogadora empleada.

[10] Accesible a través de http://www.larramendi.es/i18n/consulta_aut/busqueda.cmd. Consulta: 28/09/2015
[11] Accesible a través de http://labs.europeana.eu/data/military-maps-and-drawings-from-the-spanish-ministry-of-defence. Consulta: 28/09/2015

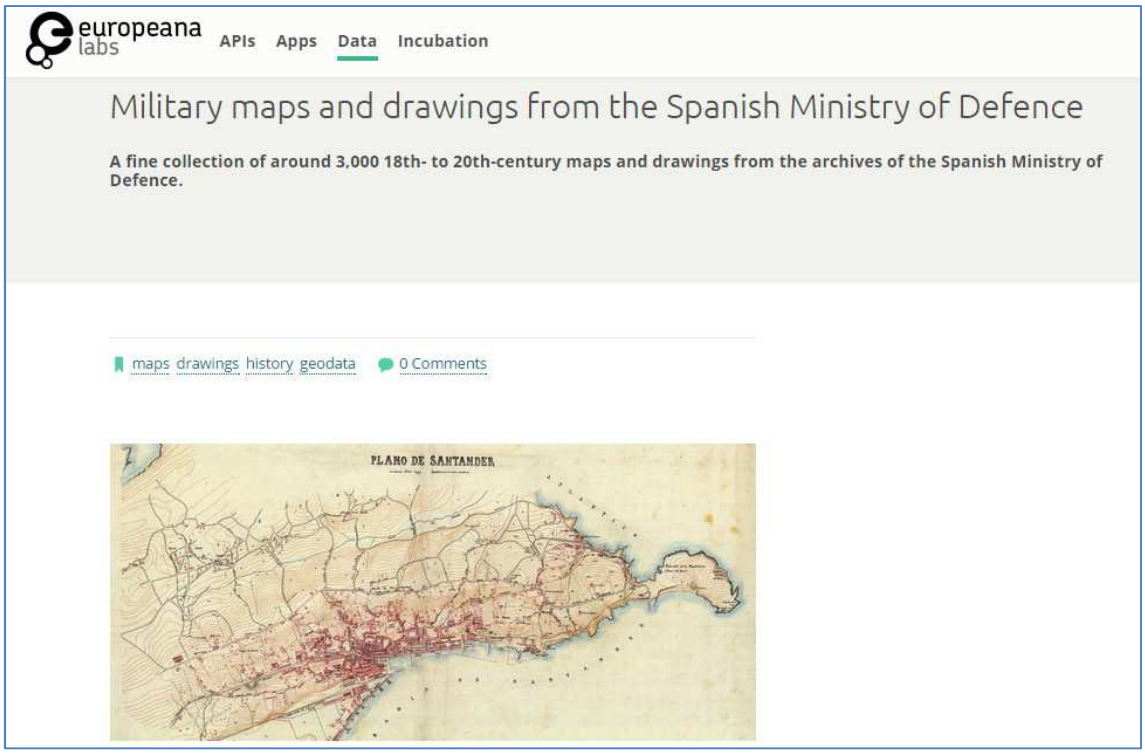

Figura 3. Biblioteca Virtual del Ministerio de Defensa en Europeana Labs.

CONCLUSIONES

El uso del modelo de datos de Europeana proporciona el marco de actuación necesario a través del cual, bibliotecas, archivos y museos pueden iniciar un proceso de semantización de sus bases de datos. El uso de este modelo no responde a una prescripción técnica aislada de un proyecto concreto sino que, por el contrario, su aplicación supone alinear las colecciones digitales en el desarrollo de las políticas de uso de tecnología Linked Open Data.

Colecciones digitales españolas como la *Biblioteca Virtual del Ministerio de Defensa*, la *Biblioteca Digital AECID* o la *Biblioteca Virtual de la Provincia de Málaga*, a través de la adecuación de este modelo de datos, gozan de un importante protagonismo dentro de la plataforma Europeana y representan un ejemplo de buenas prácticas a seguir para el desarrollo y evolución de otras colecciones digitales.

El enriquecimiento semántico es un requisito indispensable para evitar la opacidad de las colecciones digitales y aumentar la relevancia de las mismas dentro de la Web. La aplicación y uso del modelo de datos de Europeana se erige como la oportunidad ideal para incorporar un carácter semántico a colecciones que emplean un modelo de descripción plano.

BIBLIOGRAFÍA

AGENJO BULLÓN, X; HERNÁNDEZ CARRASCAL, F. (2015). "Cómo y qué consumir en el mundo Linked Open Data; cómo y qué producir en Linked Open Data". En: *XIV Jornadas Españolas de Documentación*, (Gijón 28, 29 y 30 de mayo de 2015) p. 169-195. [Disponible en: http://www.fesabid.org/sites/default/files/repositorio/actas_fesabid_2015.pdf. Consulta: 28/09/2015].

CHARLES, V.; ISAAC, A. (2015). Enhancing the Europeana Data Model (EDM). *Europeana Professional Website* [en línea]. [Disponible en: http://pro.europeana.eu/files/Europeana_Professional/Publications/EDM_WhitePaper_17062015.pdf. Consulta: 28/09/2015].

CLAYPHAM, R.; CHARLES, V.; ISAAC, A. (2014). Definition of the Europeana Data Model v5.2.6. *Europeana Professional Website* [en línea]. [Disponible en: http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM%20Definition%20v5.2.6_01032015.pdf. Consulta: 28/09/2015].

CLAYPHAM, R.; ISAAC, A. (2013). Europeana Data Model Primer. *Europeana Professional Website* [en línea]. [Disponible en: http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Primer_130714.pdf. Consulta: 29/08/2015].

COUSINS, J.; POOLE, N.;RACINE, B. (2013). Europeana Strategy 2015-2020. *Europeana Professional Website* [en línea]. [Disponible en: http://pro.europeana.eu/documents/858566/640ac847-0dfc-4b01-9f36-d98ca1212ec9. Consulta: 28/09/2015].

DIGITAL PUBLIC LIBRARY OF AMERICA (2015). DPLA Metadata Application Profile v4.0. *Digital Public Library of America Website* [en línea]. [Disponible en: http://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf. Consulta: 28/09/2015]

DIGITAL PUBLIC LIBRARY OF AMERICA (2015). Strategic Plan 2015 through 2017. *Digital Public Library of America Website* [en línea]. [Disponible en: http://dp.la/info/wp-content/uploads/2015/01/DPLA-StrategicPlan_2015-2017-Jan7.pdf. Consulta: 28/09/2015]

EUROPEANA FOUNDATION (2012). EDM Case Study: The Polymath Virtual Library and EDM. *Europeana Professional Website* [en línea]. [Disponible en: http://pro.europeana.eu/polymath-edm. Consulta: 28/09/2015].

ISAAC, A. (2015). "The Europeana Data Model: A living model 5 years on" [en línea]. En: *Europeana Blog*. 17 jun. 2015. [Disponible en: http://pro.europeana.eu/blogpost/the-europeana-data-model-a-living-model-5-years-on. Consulta: 28/09/2015].