Revista Española de Documentación Científica
41(4), octubre-diciembre 2018, e217
ISSN-L:0210-0614. https://doi.org/10.3989/redc.2018.4.1535

ESTUDIOS / *RESEARCH STUDIES*

La publicación en Linked Data de registros bibliográficos: modelo e implementación

Jose A. Senso*, Wenceslao Arroyo Machado*

*Universidad de Granada, Departamento de Información y Comunicación
Correo-e:jsenso@ugr.es | ORCID iD: https://orcid.org/0000-0002-6553-6522
Correo-e: wences@correo.ugr.es | ORCID iD: https://orcid.org/0000-0001-9437-8757

Recibido: 06-11-2017; 2ª versión: 14-02-2018; Aceptado: 16-02-2018.

Cómo citar este artículo/Citation: Senso, J. A.; Arroyo Machado, W. (2018). La publicación en Linked Data de registros bibliográficos: modelo e implementación. *Revista Española de Documentación Científica*, 41 (4): e217. https://doi.org/10.3989/redc.2018.4.1535

Resumen: Las bibliotecas se encuentran muy vinculadas a Linked Data (LD) debido al alto nivel de estructuración de sus datos, aunque los proyectos relacionados con ello son elaborados principalmente por grandes bibliotecas. En el presente trabajo se ha determinado su estado de la cuestión, analizando algunos de los proyectos referentes, ciclos de vida y herramientas que intervienen durante el proceso, estableciendo tras ello una metodología y llevando acabo su implementación al completo, convirtiendo registros bibliográficos en Linked Data, enriqueciéndolos por medio de otros conjuntos de datos y poniéndolos al alcance de todo el mundo. De este modo, se ha realizado un estudio de caso usando para ello un conjunto de registros extraídos de la Biblioteca Universitaria de Granada con el fin de conocer, de primera mano, algunos de los problemas que se puede encontrar cualquier centro que desee convertir sus registros a Linked Data sin necesidad de tener que cambiar de sistema de automatización de bibliotecas.

Palabras clave: Linked Data; Linked Open Data; Open Data; registros bibliográficos; MARC 21; Bibframe; conversión; migración.

Publishing bibliographic records on Linked Data: model and implementation

Abstract: Libraries are closely related to Linked Data (LD) due to the high level of structuring of their data, although the projects related to it are elaborated mainly by large libraries. In the present work, the state of the question has been determined, analyzing some of the referring projects, life cycles and tools that intervene during the process, establishing a methodology and carrying out its full implementation, converting bibliographic records into Linked Data, enriching them by means of other data sets and making them available to everyone. In this way, a case study has been carried out using a set of bibliographic records from the library of the University of Granada in order to know, firsthand, some of the problems that can be found by any Information Unit that wishes to convert its records to LD without having to change their library automation system.

Keywords: Linked Data; linked Open Data; Open Data; bibliographic records; MARC 21; Bibframe; conversion; migration.

1. INTRODUCCIÓN

Desde que la Biblioteca del Congreso de los Estados Unidos anunciase, en 2011, su intención de profundizar en entornos abiertos y compartidos en el trabajo bibliotecario hasta la actualidad (Marcum, 2011), las bibliotecas están viendo cómo se están transformando la mayoría de herramientas, lenguajes y sistemas que emplean para realizar su trabajo y que hasta ahora parecían inamovibles. Sin duda alguna es envidiable observar la capacidad que tiene el mundo de las bibliotecas para adaptarse a los nuevos avances si se contempla que estos pueden repercutir directamente en un mejor servicio y una ampliación en los productos que ofrecen.

Evidentemente, esta adaptación lleva un tiempo –más o menos largo- ya que estos procesos no sólo deben madurar desde el punto de vista tecnológico, sino que además deben calar en los profesionales que los deben llevar a la práctica. Pero parece claro que los conceptos de datos abiertos, Linked Data y RDF son la base (tanto filosófica como técnica) de los modelos, lenguajes y mecanismos para describir y catalogar en el futuro.

Entendemos esta última actividad, la catalogación, como el paradigma del trabajo bibliotecario (Snow, 2011; Carbonero y Dolendo, 2013). Catalogar es un proceso complejo y arduo que, si se hace de manera concienzuda, requiere de mucho esfuerzo y profesionalidad por parte de las personas encargadas de llevarlo a cabo. Sin embargo, en la actualidad, el producto de ese trabajo se guarda en catálogos que, inexorablemente, pasan a formar parte de la Internet invisible o profunda. Esto implica que sea necesario que alguien se moleste en realizar una consulta determinada en una interfaz concreta para acceder a ese registro tan bien catalogado. Por si además fuera poco, la búsqueda en esa interfaz está plagada de continuos obstáculos (control de autoridades, normalización de materias, operadores booleanos) que el usuario medio desconoce y que, posiblemente, tampoco tenga la obligación de conocer. Si bien es cierto que en la actualidad los OPACs han logrado implementar opciones muy interesantes, como las herramientas de descubrimiento, estas también evidencian problemas recurrentes (Ávila-García y otros, 2015). Además, el hecho de que ese registro entre a formar parte de un catálogo concreto –generalmente en formato propietario- impide que se pueda compartir de una forma limpia, transparente y automática, obstaculizando el intercambio de información y limitando su acceso a un número reducido de personas.

Si hace unos años los catálogos colectivos aparentaban ser una posible opción que permitiera interconectar descripciones, lo cierto es que con el paso del tiempo se ha demostrado que no son una vía válida para lograr la interoperabilidad deseada. Principalmente porque muchos de los problemas antes comentados no se terminan de solucionar, y es fácil encontrarse con una complicada gestión en el control de autoridades o de materias (Stumpf, 2003; Marais, 2009; Wolverton, 2005).

El objetivo de lograr la interoperabilidad que permita compartir datos, recursos, en definitiva, esfuerzos, pasa inexorablemente por dos principios: no abandonar la web, ya que es el entorno ideal (está normalizado, ya existen multitud de proyectos a los que es posible sumarse, el usuario se ha acostumbrado a trabajar principalmente ahí…, los motivos son infinitos) y aprovecharse de las bases tecnológicas y filosóficas que de ella emanan. Por ese motivo, un mecanismo que facilite la publicación de datos de manera normalizada y sencilla, y que permita seguir formando parte de la web, huyendo de formatos propietarios, parece que no solo es un deseo, sino una realidad. Y eso es precisamente lo que propone Linked Data.

En este sentido cabría la opción de pensar en el formato MARC como posible solución. Este formato, ahora denostado por muchos e incluso dado por muerto (Tennant, 2002; Beastall, 2016), ha servido durante muchos años como el principal mecanismo para el intercambio de registros bibliográficos en los centros de todo el mundo. Pero lo cierto es que arrastra una serie de problemas tanto de índole técnico, perfectamente descritos por Tennant (2002), entre otros, como de filosofía, que lo convierten en inviable en este nuevo entorno. Y es que una cosa es intercambiar registros y otra datos. Aunque pudiera parecer lo mismo, lo cierto es que un registro bibliográfico está constituido por innumerables datos (autor, título, lugar, editorial, fecha…) que vinculados con diferentes datos de otros datasets podrían aportar información individual de cada uno de ellos, ofreciendo al usuario un sinfín de nueva información (dónde nació ese autor, historia local de ese lugar, datos coetáneos…) y, a la biblioteca, la posibilidad de ampliar sus horizontes más allá del catálogo. Y eso es algo que con el formato MARC no se puede hacer.

Si conseguimos ofrecer los datos de las bibliotecas de manera abierta, vinculada y vinculable es posible que se puedan reutilizar, aumentando el valor de las bibliotecas, ya que tendrían mucho que ofrecer puesto que tienen mucho camino andado. No debemos olvidar que el ámbito bibliotecario, archivero y museístico está muy acostumbrado a realizar su trabajo en un entorno normalizado. El papel que pueden desempeñar este tipo de centros de información es fundamental, por el uso de

programas y la calidad del trabajo que realizan sus profesionales (Peset y otros, 2011). Los documentos y metadatos que estas instituciones tienen entre sus manos alcanzan un gran nivel de estructuración, suponiendo, en especial las bibliotecas, un terreno idóneo para iniciativas de este tipo (Sulé y otros, 2016), siendo muchos los trabajos sobre proyectos basados en esta idea (Deliot, 2014; Hallo y otros, 2014; Taylor y otros, 2013; Vila-Suero y otros, 2012), así como los que analizan el estado de la cuestión de Linked Data en el mundo de las bibliotecas (MacKenzie y otros, 2017; Torre-Bastida y otros, 2015; Peset y otros, 2011). Si miramos con una perspectiva más amplia, lo que puede aportar el uso de Linked Data está perfectamente definido tanto dentro de nuestra área, con repositorios (Subirats y otros, 2012), museos (Wang y otros, 2008) y archivos (Hidalgo-Delgado y otros, 2016), como en otras disciplinas como la educación, la medicina y un largo etcétera.

Emplear este mecanismo para vincular los datos que aparecen en los registros bibliográficos con otros datasets con el fin de interconectar información aporta, tanto a los usuarios como a las bibliotecas, una mejora en la visibilidad (del dato y de la institución que lo ofrece), permite reaprovechar los datos de los registros publicados y añadir nuevos, establece vínculos con otros servicios y favorece el desarrollo de mashups, además de facilitar el modelado de "cosas de interés" relacionadas con un recurso bibliográfico, como personas, lugares, eventos y temas. Y todo eso sin afectar a los modelos de la fuente de datos.

Los principios sobre los que se sustenta el sistema de publicación denominado Linked Data se establecieron en 2010 (Berners-Lee, 2010) sobre un mecanismo que aporta hasta un máximo de 5 estrellas en función a cómo se compartan los datos. Una estrella se asigna si tan solo se publican los datos con licencia abierta, independientemente del formato; las dos estrellas las tienen los conjuntos de datos (también llamados datasets) que se publiquen como datos estructurados, aunque fueran propietarios; tres estrellas significa que se emplean formatos no propietarios; se añade una estrella más si a ese dataset se le incluyen URIs que permiten identificar y apuntar hacia esos datos, y las cinco estrellas se consiguen si los datos que se ofrecen ya están enlazados a otros, con el fin de que tengan un contexto más claro. En 2014 surgió una propuesta que ampliaba hasta siete estrellas este método de puntuar la calidad de los datos que se comparten (Hyvönen y otros, 2014), teniendo en cuenta si en el dataset se añadían vocabularios y si, además, se valoraban otros criterios relacionados con los datos aportados.

La importancia de Linked Data dentro del ámbito de las bibliotecas se incrementó en 2004, cuando el Consorcio WWW recomendó que éstas publicasen sus datos utilizando tecnologías de la Web Semántica para incrementar su impacto digital y utilidad social (Hallo y otros, 2016). En 2010 surge el W3C Library Linked Data Incubator Group para "ayudar a aumentar la interoperabilidad global de datos de las bibliotecas en la web", que concluyó un año más tarde (Bermès y otros, 2011), coincidiendo con el anuncio de la Biblioteca del Congreso de Bibframe (Bibliographic Framework), planteado como la evolución del formato MARC 21 a la Web Semántica y el Linked Data (Kroeger, 2013).

Desde entonces, cada vez son más los proyectos realizados en el entorno de las bibliotecas con Linked Data como principal protagonista. La Biblioteca Nacional de España, la British Library, La Bibliothèque National de Francia, Europeana o la propia Biblioteca del Congreso son constantes que aparecen en todos los estados de la cuestión (Hallo y otros, 2016; Papadakis y otros, 2015; Torre-Bastida y otros, 2015; Wenz, 2013; Peset y otros, 2011), así como en estudios de caso (Deliot, 2014; Hallo y otros, 2014; Vila-Suero y otros, 2012; Wenz, 2013).

La mayoría de estos proyectos tienen dos constantes. Por un lado, son llevados a cabo por grandes instituciones que, en muchas ocasiones, han necesitado de la ayuda de un tercero (empresa o universidad) para finalizar con éxito sus implementaciones. Por otro, no se puede observar una metodología clara y uniforme para la transformación de los registros bibliográficos a Linked Data.

Los motivos que justifican la primera constante parecen evidentes: la mayoría de instituciones hasta ahora mencionadas carecen del potencial (económico, tecnológico y/o humano) para realizar una tarea tan especializada y que se adentra tan claramente dentro del entorno informático. La segunda constante requiere tener en cuenta más ítems a valorar, y todos ellos se pueden aglutinar en las fuentes de datos. De su calidad, licencias, vocabularios y ontologías empleados, datasets usados para su enriquecimiento a través del enlazado, su método de publicación y las tecnologías empleadas durante este proceso –entre otros aspectos- dependen variables que pueden hacer cambiar el transcurrir de un proyecto concreto. Todo esto hace que puedan ser múltiples los caminos a elegir, impidiendo establecer una única metodología.

Tantas opciones ofrecen ciclos de vida diferentes, que se suelen definir sobre la base de objetivos y necesidades a cubrir. No obstante en la mayoría de los casos es posible encontrar un común deno-

minador, que viene determinado por aquellas fases que se repiten, y que podríamos entender que forman parte del ciclo de vida común a la mayoría de proyectos Linked Data y Linked Open Data que tengan como objetivo tanto la publicación de datos como su posterior enriquecimiento.

En un entorno ideal, lo lógico sería que el propio programa de automatización de bibliotecas fuera capaz de publicar los datos catalogados en Linked Data. De esa manera, este proceso sería totalmente transparente y actualizado conforme se pone al día el catálogo de la biblioteca con sus incorporaciones, eliminaciones, etc. Es más, lo ideal sería que en el mismo proceso de catalogación se pudiera escoger qué datos del registro con el que se está trabajo son susceptibles de ser vinculados con otros datasets, ya precargados en el sistema, y que desde él se realizara ese vínculo. Sin embargo, en la actualidad esto no sucede, ya que son muy pocos los sistemas integrados que permiten realizar este tipo de tareas o similares. El mercado ofrece pocas soluciones, entre las que destacan la española DigiBIB (con variantes para archivos, DigiArch, y para museos, DigiMus), de la empresa Digibis, y el servicio Innovative Linked Data de la norteamericana Innovative Interfaces Inc.; que se ofrece como un extra a sus programas Sierra y Polaris. Posiblemente en los nuevos requisitos funcionales para este tipo de programas deberían incluirse varios ítems que valoraran positivamente aquel software que fuera capaz de realizar estas funciones.

Esto nos lleva a un escenario poco homogéneo, que obliga a procesar todos los registros bibliográficos en un hábitat diferente del programa de automatización, y en entornos tan diferentes como bibliotecas, datasets y sistemas de gestión bibliotecaria existan. De esa forma es fácil entender que no se pueda contar con una metodología única y clara que pueda adaptase, de manera flexible, a todas las bibliotecas que quieran compartir sus datos a través de Linked Data. Esa es la principal motivación de este trabajo: establecer un modelo que ayude a las bibliotecas a definir un flujo de trabajo que facilite el proceso de publicar en Linked Data los registros bibliográficos que almacenan en sus catálogos automatizados. Por lo tanto establecemos como objetivo principal el elaborar una metodología que sirva para convertir registros disponibles en cualquier biblioteca, y en cualquier formato, en Linked Data, enriqueciéndolos por medio de otros conjuntos de datos con el fin de ponerlos a disposición de la comunidad.

Tras la elaboración de dicho modelo de transformación se procederá a su implementación en un conjunto de datos pequeño, con el fin de observar las posibles deficiencias que este método pueda tener, así como determinar las principales dificultades que conlleva este proceso de transformación.

2. MATERIALES Y MÉTODOS

Teniendo en cuenta el doble objetivo del trabajo, por un lado, crear un modelo de transformación de registros bibliográficos, y, por otro, la creación de un piloto que permita averiguar si ese modelo es factible, es necesario aplicar diferentes metodologías. Así, para la primera fase, se procederá a un estudio bibliográfico con el fin de determinar el estado de la cuestión de Linked Data en bibliotecas, prestando especial atención a la forma en la que han procedido los principales proyectos a nivel internacional.

Una vez realizada esa fase se obtendrá tanto una visión global de cómo se han desarrollado dichos proyectos como los elementos necesarios para realizar una propuesta metodológica. Los autores son plenamente conscientes de que uno de los factores que determinan, en mayor o menor manera, el éxito de una metodología se encuentra en que existan suficientes herramientas en el mercado que faciliten su implementación. Por ese motivo entendemos necesario realizar un estudio de las principales aplicaciones que se puedan emplear en cada una de las etapas del ciclo de vida de ese conjunto de datos, con el fin de saber si esta propuesta puede ser asumible por cualquier institución que desee aplicarla.

Una vez diseñado el modelo de conversión de registros, se procederá a realizar una pequeña implementación, a modo de piloto, que permita determinar tanto la viabilidad de la propuesta como los principales problemas encontrados durante su puesta en marcha, así como las posibles soluciones que se puedan plantear.

Con el fin de trabajar dentro de un entorno lo más cercano posible a la realidad, y conocer así las dificultades más comunes a las que se pueda enfrentar cualquier centro que desee realizar este proceso, se han empleado un conjunto de registros bibliográficos procedentes de la Biblioteca General de la Universidad de Granada, empleando para el trabajo inicial todos los formatos de exportación que ofrece el software de automatización que allí se emplea. De esa manera, además, se obtendrá información fidedigna sobre cuál es el mejor punto de partida del dataset. Dado que no se trata de un trabajo exhaustivo, se ha optado por escoger los datos de autor, título, publicación, materia e ISBN de cada registro. Los autores entienden que este subconjunto aporta la información necesaria como para determinar la idoneidad del método, además

de suministrar los datos necesarios que se requieren para conocer la fiabilidad del sistema. Emplear otros campos no añadiría información complementaria a la que aporten los escogidos.

Dada la gran cantidad de registros con los que cuenta esta biblioteca, se optó por trabajar con una muestra (nunca inferior a los 1.000 registros) y centrados en una única temática. De esa manera sería más sencillo apreciar las posibles desviaciones que se produjeran durante el proceso de conversión. Dado que en esta fase del trabajo no se ha realizado aún ningún estudio sobre las herramientas disponibles, los autores entienden que no deberían de tener predisposición alguna a este respecto, por lo que salvo el uso de aplicaciones específicas para el trabajo con ficheros bibliográficos (MarcEdit o MARC Editor), será la fase de evaluación de herramientas la que determinara el conjunto de ellas con las que llevar a cabo el proceso de transformación de registros.

3. DISEÑO DEL MODELO

Como ya se ha comentado con anterioridad, la mayoría de proyectos de ámbito internacional se han centrado en el ámbito de las grandes bibliotecas. Y aunque es cierto que existen aportaciones de otro tipo de instituciones, tal y como lo demuestra el trabajo del Library Linked Data Incubator Group (Isaac y otros, 2011), la mayoría de ellos no aportan la gran cantidad de información adicional que sí ofrecen los Centros Nacionales y que facilitan mucho la labor de conocer cómo se ha realizado el proceso de conversión.

Si se analizan las diferentes aportaciones enfocadas al análisis de la situación actual de Linked Data en bibliotecas (Hallo y otros, 2016; Papadakis y otros, 2015; Torre-Bastida y otros, 2015; Peset y otros, 2011), se observa que todas ellas coinciden al hablar de un conjunto de proyectos que se pueden considerar paradigmáticos dentro de este entorno. De esta manera, un análisis de los procesos llevados a cabo en la Biblioteca Nacional de España, la Brisith Library, la Bibliothèque Nationale de France, Europeana y la Library of Congress aporta el conocimiento necesario para saber cuáles son las metodologías más empleadas en la actualidad para llevar a cabo la conversión de registros bibliográficos a Linked Data. Entre las principales características de dichos proyectos destacamos:

- La Biblioteca Nacional de España ofrece, a través de su portal (Biblioteca Nacional de España, 2016a), acceso al catálogo bibliográfico y de autoridades en Linked Open Data. Para ello se han transformado registros desde MARC21 a RDF por medio de un proceso automatizado con el software Marimba (Vila-Suero y Gómez-Pérez, 2013), permitiendo el descubrimiento de enlaces hacia otros datasets por medio de otros programas, como Silk (Volz y otros, 2009). La publicación y consulta de sus datos es posible gracias al repositorio RDF Virtuoso (OpenLink, 2015) y la interfaz Pubby (Cyganiak y Bizer, 2011). El ciclo de vida de los registros está compuesto por siete pasos, donde destaca principalmente la fase de limpieza de datos y el desarrollo de aplicaciones (Vila-Suero y otros, 2012).
- La British Library (BL) cuenta con la British National Bibliography en Linked Open Data (The British Library, 2014). Los registros no los transforma de MARC21 a RDF, sino que primero identifica "objetos de interes" (incluyendo conceptos y abstracciones) y los declara por medio de URIs propias. Tras eso, se describen las clases y sus relaciones entre sí, para lo cual definieron sus propias clases y propiedades, documentadas en el British Library Terms RDF Schema (Deliot, 2014). Los datos enlazados de la BL siguen dos modelos diferenciados, ya que uno es para libros (British Library data model for books) y otro para publicaciones seriadas (Brisith Library data model for serials).
- La Bibliothèque Nationale de France ha trabajado con diferentes bases de datos vinculando metadados de documentos en papel con su versión digitalizada. El producto final se puede visualizar desde su portal de datos (Bibliothèque National de France, 2014). Aquellas bases de datos que eran no interoperables las han transformado en datos estructurados e intercambiables empleando principalmente RDF. A los recursos que han ido generando se les ha asignado un identificador permanente denominado ARK (Archival Resource Key) (Wenz, 2013).
- La reciente remodelación de la página de datos de Europeana (Europeana, 2017) permite acceder a gran cantidad de información específica, tanto de su modelo de datos EDM (Europeana Data Model), como de los procesos llevados a cabo para su transformación en Linked Data. EDM incluye conexiones a fuentes externas y reutiliza elementos procedentes de vocabularios ya establecidos,como Dublin Core, OAI-ORE, Skos y CIDOC-CRM. Los datos que se pueden encontrar para cada clase de recurso son la propia representación del objeto, sus datos descriptivos, los datos referentes al proveedor y los metadatos descriptivos asignados tanto por éste como por Europeana.

- Posiblemente, el proyecto que más vínculos establece con otros datasets sea el capitaneado por la Biblioteca del Congreso (Library of Congress, 2016b). En realidad, el hecho de poder contar con Bibframe como modelo de datos de descripción bibliográfica facilitó mucho el proceso de convertir los registros MARC 21 de los que disponía esta biblioteca (Library of Congress, 2012).

Tras el análisis exhaustivo de los pasos que han seguido estas bibliotecas, junto al estudio de la hoja de ruta establecida en el proyecto BIBLOW (MacKenzie y otros, 2017) y considerando los métodos empleados en otros trabajos (Hallo y otros, 2014), consideramos que tenemos la información suficiente como para proponer un modelo uniforme que permita realizar la transformación de registros bibliográficos a Linked Data independientemente del software o del entorno en el que estos se encuentren. Dicho modelo se muestra en la tabla I.

Al mismo tiempo se ha obtenido una visión global de los programas más utilizados y que se podrían emplear en el desarrollo y puesta en marcha de cada una de esas etapas. Sin el ánimo de ser un listado exhaustivo, nos hemos centrado en la agrupación de herramientas open source o gratuitas que más se emplean en la actualidad. De esa forma aumenta la posibilidad de que este modelo pueda ser llevado a la práctica. La tabla II muestra dicho listado.

4. CASO DE ESTUDIO

Con el fin de conocer si las fases propuestas son viables se ha realizado un estudio piloto sobre un conjunto de registros bibliográficos. A continuación se muestra su desarrollo dentro de cada una de las 6 fases del modelo propuesto.

4.1. Determinar los datos

Teniendo en cuenta que la mayoría de bibliotecas que desearan emplear este método contarían con registros ya creados, se partió de la idea de trabajar con un conjunto de datos homogéneo, y por ello se optó por extraerlos de la misma fuente. Por ese motivo se emplearon los almacenados en el Catálogo General de la Biblioteca de la Universidad de Granada (BUGR en adelante). También se consideró que no era necesario trabajar con todos los campos que ofrecía cada registro. Primero, porque el objetivo de probar el modelo se podía llevar a cabo con un conjunto de datos suficientemente representativo; segundo, porque se partía del hecho de que los problemas significativos aparecerían al gestionar esos campos, y los siguientes serían meras repeticiones; y tercero, porque de esa manera se facilitaba la etapa de limpieza de datos, lo que agilizaría su posterior carga en el sistema. Por esos motivos se optó por emplear los datos de autor, título, publicación, materias e ISBN. En una primera fase de este es-

Tabla I. Propuesta de metodología para publicar registros bibliográficos como Linked Data

Etapa	***Descripción***	***Tareas***
1. Determinar	Identificación y descripción de los datos	a. Identificar y analizar los datos y fuente de datos (software, formato, base de datos…)
		b. Identificar su licencia
		c. Determinar una licencia
2. Limpiar	Almacenamiento y corrección de los datos	a. Data curation
3. Modelar	Desarrollo de un vocabulario para describir los datos en formato RDF	a. Seleccionar los vocabularios
		b. Creación de mapa
		c. Asignar URIs
4. Generar	Generación de los recursos RDF	d. Seleccionar las tecnologías para la generación de RDF
		e. Transformar los datos fuente en RDF
		f. Validarlo
5. Enlazar	Conectar el dataset a otros que lo enriquezca	a. Buscar datasets relevantes
		b. Descubrir relaciones
		c. Enlazar
		d. Verificar los enlaces
6. Publicar	Publicación del dataset	a. Escoger el formato y plataforma
		b. Publicar el dataset
		c. Publicar sus metadatos

Tabla II. Herramientas más empleadas en cada uno de los procesos

Almacenamiento y gestión de datos				
Nombre	URL	Descripción	Licencia	Plataforma
Apache Hadopp	http://hadoop.apache.org/	Framework de software open-source para el almacenamiento distribuido de conjuntos de datos muy grandes en clusters de ordenadores.	Apache License 2.0	Multiplataforma
Cloudera Distributed Hadoop (CDH)	http://www.cloudera.com	Distribución de Apache orientada al mundo empresarial	Apache License 2.0	Linux
MongoDB	https://www.mongodb.com	Base de datos NoSQL	GNU AGPL 3.0	Multiplataforma
Extracción y limpieza de datos				
Spoon - Pentaho's Data Integration	http://community.pentaho.com/projects/data-integration/	Herramienta open-source para la extracción, transformación, transporte y carga de datos (ETL)	Apache License 2.0	Multiplataforma
Virtuoso Sponger	https://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtSponger	Se trata de un componente middleware de Virtuoso Open-Source (VOS) que permite importar datos en diversos formatos (CSV, RSS, vCard…) y transformarlos en RDF	GNU General Public License 2.0	Multiplataforma
D2RQ	http://d2rq.org	Sistema open-source que permite acceder a bases de datos relacionales como grafos RDF virtuales, pudiendo lanzar consultas SPARQL en bases de datos no RDF, así como exportar la base de datos en RDF	Apache License 2.0	Multiplataforma
OpenRefine	http://openrefine.org	Herramienta ETL (Extraer, Transformar y Cargar) enfocada a la limpieza, transformación, exploración y enlazado de datos procedentes de diversos formatos. Sus funciones se pueden expandir con el uso de extensiones, destacando RDF Refine extension o DBpedia extension	BSD	Multiplataforma
GraphDB Free Edition	http://ontotext.com	Se trata de un repositorio semántico, un sistema de base de datos NoSQL que permite almacenar, consultar y gestionar datos estructurados. Utiliza ontologías para razonar automáticamente sobre los datos.	Licencia libre tipo RDBMS	Multiplataforma

Modelización				
Protégé	http://protege.stanford.edu/	Herramienta open-source que permite la construcción de modelos de dominio y aplicaciones basadas en el conocimiento con ontologías. Cuenta con una versión web y otra de escritorio. Es compatible con la última versión del Lenguaje de ontologías OWL 2 y especificaciones RDF de la World Wide Web Consortium (W3C)	FreeBSD	Multiplataforma
CmapTools Ontology Editor (COE)	http://cmap.ihmc.us/coe/test/v401ReleaseNotes.html#	Versión de CmapTools, herramienta para los mapas conceptuales, orientada a construir, compartir y visualizar ontologías OWL	-	Multiplataforma
OntoWiki	http://ontowiki.net/	Esta herramienta open-source permite la edición del contenido de archivos RDF de una forma muy visual, del mismo modo que un editor WYSIWIG para documentos de texto.	GNU General Public License 2.0	Multiplataforma
OOPS!	http://oops.linkeddata.es/	Se trata de una herramienta online de validación que permite detectar algunos de los errores más comunes que aparecen al desarrollar ontologías	-	Online
W3C RDF Validation Service	https://www.w3.org/RDF/Validator/	Herramienta online de W3C para la validación y visualización de documentos RDF (RDF/XML).	-	Online
Enlazado				
Limes	http://aksw.org/Projects/LIMES.html	Framework que implementa métodos eficientes en tiempo para el descubrimiento de enlaces a gran escala basados en las características de los espacios métricos	GNU General Public License	Multiplataforma
Silk	http://silkframework.org/	Open-source framework para combinar fuentes de datos heterogéneas, permitiendo generar enlaces entre elementos de datos contenidos en distintas fuentes	Apache License 2.0	Multiplataforma
Publicación				
Virtuoso	https://virtuoso.openlinksw.com/	Servidor multiplataforma escalable para el acceso a datos, integración y gestión de bases de datos relacionales, RDF y XML con un servidor de aplicaciones, de servicios Web	Apache License 2.0	Multiplataforma

tudio se planteó la importancia que tendría, de cara a establecer posteriores relaciones, el empleo de los campos 76X-78X (campos de enlace), sin embargo, y para que esta información fuera relevante, era necesario que un elevado número de registros contaran con esta información. En este caso en concreto, menos del 1% de registros cumplimentaban esos campos, por lo que carecía de sentido su uso. Algo similar sucede con los ejemplares. Para que su empleo aporte algo desde el punto de vista semántico es preciso contar con suficiente información. En nuestro caso tan solo disponíamos del campo 945, con la signatura topográfica como único dato. A continuación se realizó, de manera consecutiva, la descarga de registros para proceder, después, a la evaluación de la información, la selección de campos de interés y la extracción de datos.

Para su descarga, la BUGR dispone de una base de datos en la que se encuentran almacenados todos los registros bibliográficos, pudiendo acceder a ellos a través de dos catálogos web, por medio de dos motores de búsqueda Adrastea (http://adrastea.ugr.es) y VELETA (http://bencore.ugr.es), o vía servidor Z39.50. No obstante, el hecho de que este último servidor limitara la cantidad de registros recuperados a 500 obligó a que se tuviera que desestimar su uso.

Con el fin de conocer mejor las limitaciones del modelo se decidió emplear un conjunto de registros que fuese de temática familiar a estos autores, por lo que se optó por recopilar todos los que tuvieran asignadas las materias "Biblioteconomía" o "Documentación". De esa manera, y al contar con información a priori de la fuente con la que se trabaja (conocimiento de las series, los autores, las editoriales, etc.), se podría tener un mayor control sobre los resultados y tener criterio suficiente para conocer si los posibles errores futuros se debían al dato o a la aplicación del método. En total se recuperaron 1.251 registros. Estos registros se descargaron empleando el OPAC tradicional, Adrastea, al ser el único que permite realizar este proceso de forma masiva y, además, exportar a diferentes formatos (pantalla completa, presentación abreviada, ProCite, Endnote-Refworks y MARC) todos ellos como un archivo de texto plano.

Para decidir con qué formato se trabajaría en las siguientes fases se realizó un análisis de cómo estaban construido cada fichero, buscando aquél que no solo facilitara la información lo más clara posible, sino que además permitiera la posterior gestión de los datos. De esa manera se buscaba un formato lo más parecido a CSV (Comma-Separated Values), TSV (Tab-Separated Values) u otro con una delimitación similar que permita separar los registros por filas y los campos por columnas.

En consecuencia, el formato escogido fue MARC por contener toda la información bruta existente acerca de todos esos registros y de la manera más normalizada y limpia posible, así como por ser un formato que, gracias al programa MarcEdit 6, podía fácilmente transformarse en CSV con aquellos campos y subcampos que quisiera.

Antes de proceder a su extracción, y también a través de la herramienta anteriormente citada, se elaboró un informe en el que se observó la frecuencia de aparición de los diferentes campos y subcampos de MARC 21 (Biblioteca Nacional de España, 2016b) para determinar si se contaba con una cantidad mínima de información en aquellos campos con los que se pensaba trabajar de cada registro (autor, título, publicación, materias e ISBN). Al mismo tiempo, y en función de la cantidad de información de la que se disponía, este proceso sirvió para determinar qué campos 6XX (encabezamiento de materia) se emplearían. En consecuencia, se estableció el uso del campo 650 (punto de acceso adicional de materia - término de materia) por aparecer en el 99% de los registros y ser términos controlados, frente al siguiente con mayor frecuencia que era el campo 655 (término de indización - género/forma) con presencia en un 52% de ellos.

El uso de este formato permitió que se pudiera trabajar con algunos campos que, en un principio, no se habían contemplado. Así, la cabecera de cada registro, los campos 001 (número de control) y 008 (códigos de información de longitud fija) aportaban información que podría permitir agilizar procesos posteriores. En algunos casos fue necesario realizar alguna modificación en los ficheros originales ya que, por ejemplo, el campo 001 sólo estaba presente en 195 registros y, además, no con valores consecutivos. Se optó por eliminarlos todos y se generaron de nuevo, enumerándolos desde el 1 hasta al 1.251. Para finalizar, se exportaron todos los registros delimitados por tabuladores a través de MarcEdit 6, seleccionando los diferentes campos y subcampos (tabla III) para exportarlos en CSV, fijando como delimitador de campos la coma, como delimitador dentro de campo el punto y coma (;) y delimitador contextual la línea vertical (|).

La última de las tareas relacionadas con esta se centraba en la identificación de la licencia con la que contaban los registros obtenidos y, en función de ella, determinar la que se emplearía en adelante. Lamentablemente, ni el sitio web de la BUGR ni las personas consultadas aportaron información al respecto.

Tabla III. Campos y subcampos de MARC 21 escogidos para su exportación en formato CSV

Información	Campo	Subcampo	Registros
Autor	100 – Punto de acceso principal-Nombre de persona	$a – Nombre de persona	624
		$d – Fechas asociadas al nombre	52
Título	245 – Mención de título	$a – Título	1.251
		$b – Resto del título	583
		$c – Mención de responsabilidad, etc.	653
Publicación	260 – Publicación, distribución, etc.	$a – Lugar de publicación, distribución, etc.	1.243
		$b – Nombre del editor, distribuidor, etc.	1.251
		$c – Fecha de publicación, distribución, etc.	1.183
Materia	650 – Punto de acceso adicional de materia -Término de materia	$a – Término de materia	1.244
		$x – Subdivisión de materia general	195
ISBN	020 – ISBN	$a – ISBN	1.015
Cabecera	Cabecera		1.251
Número	001 – Número de control		1.251
Información	008 – Códigos de información de longitud fija		1.251

4.2. Limpieza de datos

Como se puede apreciar en la tabla II, existen varias herramientas que facilitan este proceso. Tras un estudio de las posibilidades de todas ellas nos decantamos por emplear GraphDB, en su versión gratuita 8.0.1. Es una aplicación muy similar a OpenRefine, pero con opciones muy interesantes, como el poder lanzar consultas sobre los datos almacenados o externos mediante SPARQL. Esta característica fue determinante para que se optara por ella, ya que facilitaría mucho llevar a cabo alguna de las fases posteriores del modelo.

Se procedió a la importación del archivo creado en el proceso anterior, obteniendo una fila para cada registro bibliográfico y una columna para cada campo y subcampo. El objetivo que se persiguió en esta fase fue corregir al máximo los posibles errores que tuvieran los registros, con el fin de obtener "datos limpios y de alta calidad" (Montalvillo Mendizábal, 2012).

Sin duda alguna esta es la fase donde se encuentra el grueso del trabajo, la que más esfuerzos y tiempo requiere pero, al mismo tiempo, la que determinará si el resto del proceso de conversión de registros culmina correctamente o no. Del éxito de esta etapa se encuentra el poder desarrollar el resto, en especial la de enlazado, donde al cruzar los datos con los de otros datasets es crucial que estos se encuentren en el mejor estado posible, de cara tanto a poder establecer relaciones como garantizar que otros puedan consumirlos de misma manera.

Para ello, en primer lugar, se fueron revisando cada una de las columnas a través de la opción de búsqueda facetada que ofrece el GraphDB, observando rápidamente la existencia y frecuencia de fallos, como variaciones entre términos que deberían aparecer recogidos bajo una misma faceta pero que no lo hacen.

Una vez localizados, se realizaron las diferentes modificaciones a través de funciones en GREL (General Refine Expression Language) aplicables a las cadenas de texto de todas las filas de una columnas (Morris, 2015), usando además expresiones regulares con la sintaxis de JAVA, además de agrupaciones por clusters. En definitiva, se ha llevado a cabo una modificación de la manera lo más amplia posible debido a la considerable cantidad de datos. En este proceso se ha observado cómo la mayoría de registros suelen repetir los mismos errores en los mismos campos, por lo que la búsqueda de esos patrones ha sido clave para automatizar el proceso de corrección. En ocasiones estos problemas eran debidos a una mala catalogación y, en otros, por problemas ocasionados en la conversión de registros de la propia BUGR.

A decir verdad, la mayoría de errores estaban relacionados con el control de autoridades (lo que impediría que los registros resultantes se pudieran

conectar con, por ejemplo, VIAF), los puntos de acceso, los encabezamientos de materia (impidiendo su vinculación con la Lista de Encabezamientos de Materia para Bibliotecas Públicas) y la ausencia de guiones en el ISBN, elemento que también se suele emplear para el proceso de vinculación con otros datasets. En definitiva, se tuvieron que realizar una gran cantidad de transformaciones con el fin de dejar en cada celda el dato en bruto y lo más normalizado posible. La tabla IV muestra un ejemplo de cómo se ha transformado un registro y cómo queda el resultado después de este proceso.

A los autores les llamó poderosamente la atención la gran cantidad de datos con mala calidad que se obtuvieron de la primera etapa del modelo, por lo que se realizó un proceso similar pero con una muestra de datos procedentes de la Biblioteca Nacional de España (en adelante BNE). La idea era conocer si este tipo de problemas son producto del proceso de exportación de registros desde el OPAC o si, por el contrario, se deben a cuestiones propias de cada biblioteca. Para eso se realizó un proceso similar, capturando en formato MARC los registros sobre las materias "Biblioteconomía" o "Documentación" y realizando los pasos anteriormente citados. El resultado de este pequeño estudio permitió determinar que, si bien es cierto que los registros de la BNE también presentaban problemas de calidad en sus datos (especialmente en el control de autoridades), el acceso y estructura de lo extraído presenta una información más estructurada y homogénea que los de la BUGR, por lo que nos decantamos por pensar que los problemas no son principalmente de índole técnico o achacables al sistema de automatización allí empleado.

4.3. Modelación de datos

Entendemos el proceso de modelación como la selección del conjunto de herramientas conceptuales que permiten describir los datos, sus relaciones, significado y restricciones de cualquier tipo. Siguiendo el modelo propuesto, se trató de localizar aquellos vocabularios que permitieran, con la mayor precisión, realizar este proceso. Para ello se emplearon los buscadores vocab (Ontology Engineering Group, 2017) y Linked Open Vocabularies (LOV, 2016).

Tabla IV. Ejemplo de transformación de datos

Col.	***Valor original***		***Resultado***
000	00526nam a2200205 i 4500		Nuevo
			Textual
			mono
001	1		1
008	090512s1930\\\\uk\\\\\\\\\\\\000\\|\\ eng\d		eng
100$a	Kenyon, Frederic George,		Kenyon, Frederic George
			Frederic George Kenyon
100$d	1863-1952.		1863-1952
245$a	LIS education in developing countries :		LIS education in developing countries
245$b	the road ahead /		the road ahead
245$c	edited on behalf of IFLA by Ismail Abdullahi, A.Y. Asundi and C.R. Karisiddappa		edited on behalf of IFLA by Ismail Abdullahi, A.Y. Asundi and C.R. Karisiddappa
260$a	[Jaén] :		Jaén
260$b	[Mundaneum],		Mundaneum
260$c	c2010.		2010
650	\4$aUnión Europea.;\4$aDocumentación$xPublicaciones periódicas.		Unión Europea
			Documentación—Publicaciones periódicas
			Documentación
			Publicaciones periódicas
020$a	1843340534 (paperback);1843340542 (hardback)		1843340534
			1843340542

El principal vocabulario escogido para convertir los datos obtenidos de la anterior fase en Linked Data fue Bibframe 2.0 (Library of Congress, 2017b). El hecho de contar con el respaldo de la Biblioteca del Congreso de los Estados Unidos y de que, a pesar de ser un proyecto de ámbito internacional, apenas esté documentado para este tipo de procesos, parecieron motivos de suficiente peso para su elección. Por otra parte, entendemos que no es práctica la idea de construir un modelo específico, ni para esta metodología en concreto ni para cada biblioteca en particular. Por ese motivo tiene mucho sentido el empleo de Bibframe. De hecho, si las bibliotecas nacionales que se describen en el epígrafe 3 hubieran contando con un modelo de datos de Bibframe posiblemente no tendrían que haber desarrollado sus propias metodologías. Para complementarlo se emplearon también LC Bibframe 2.0 Vocabulary Extension (Library of Congress, 2016a) y MADS/RDF Primer (Library of Congress, 2015). Una vez establecidos los vocabularios a usar se desarrolló un mapa o red de las diferentes clases y subclases, propiedades y relaciones que se van a establecer con los datos obtenidos (Dimou y otros, 2016), tal y como muestra la figura 1. Para ello se emplearon las especificaciones fijadas por la propia Library of Congress para transformar MARC 21 en Bibframe (Library of Congress, 2017a).

A continuación se muestran los namespaces empleados para la construcción de dicha red:

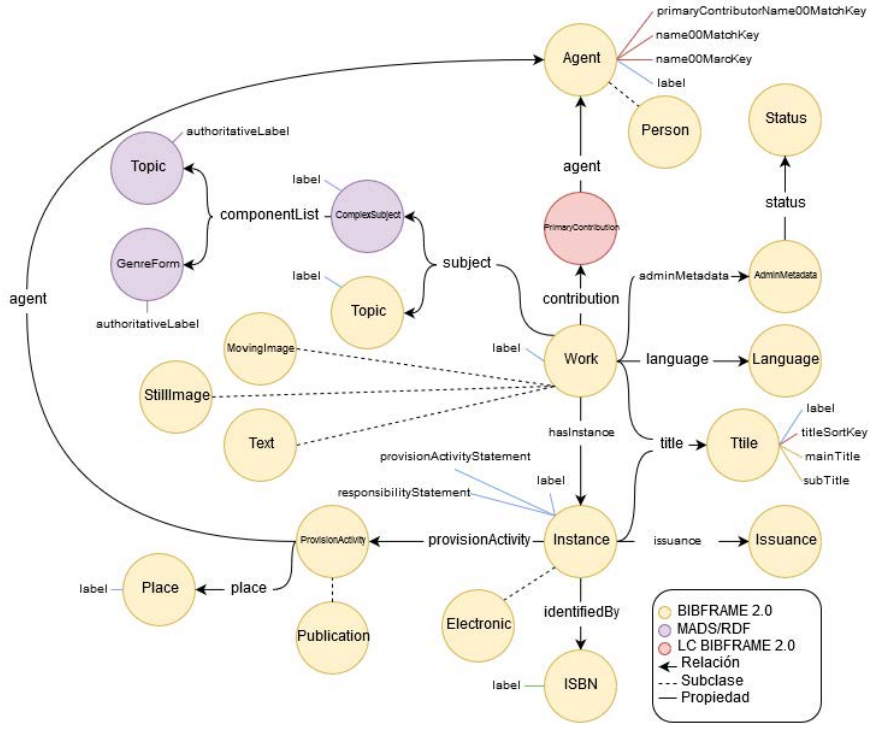
- bf: <http://id.loc.gov/ontologies/bibframe/>
- madsrdf: <http://www.loc.gov/mads/rdf/v1#>
- bflc: <http://id.loc.gov/ontologies/bflc/>
- rdfs: <http://www.w3.org/2000/01/rdf-schema#>
- rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#

Como URI base para el modelado se empleó <http://example.org/>, asignándose las siguientes para estas entidades:

- Work – <http://example.org/numero-de-control#Work>
- Instance – <http://example.org/numero-de-control#Instance>
- Topic – <http://example.org/#Topic650-nombre-de-materia>
- ComplexSubject – <http://example.org/#Topic650-nombre-de-materia>
- Person – http://example.org/nombre-de-autor

Aparte de estas URIs, también se construyeron otras dos más con el código de idioma y el nivel bibliográfico, usando con ello dos esquemas de la Library of Congress:

Figura 1. Mapa del vocabulario



- Language – <http://id.loc.gov/vocabulary/languages.html>
- Issuance – http://id.loc.gov/vocabulary/issuance.html

4.4. Generación de datos

Para esa fase se empleó el programa GraphDB, creando un repositorio que almacenase los datos RDF y, por medio del lenguaje de consultas SPARQL (Harris y Seaborne, 2013), se fueron elaborando y ejecutando las diferentes consultas para crear los grafos con Bibframe 2.0.

De este modo se ejecutaron las diferentes consultas en SPARQL, que recogen los datos anteriormente limpiados del CSV, buscando por filas y columnas, y transformándolos en grafos. Esta etapa no lleva mucho tiempo, especialmente si se compara con la de limpieza de datos, pero sí es posiblemente la más compleja, ya que hay que realizar búsquedas bastante sofisticadas y que emplean múltiples variables. Una vez finalizado este proceso se obtuvieron las marcadas 22 clases y subclases, tal y como se muestra en la figura 2, coincidiendo el contenido de cada una de ellas con las instancias previstas.

Una vez construidos los grafos se procedió a la tarea de validación, para la que se empleó la herramienta IDLab Turtle Validator (Internet & Data Lab, 2016). Para ello se realizó la exportación de todos los grafos almacenados en GraphDB empleando el formato de serialización Turtle, tras lo cual se copió el contenido de ese archivo en el validador, incluyendo tanto los namespaces como las tripletas.

4.5. Enlazado de datos

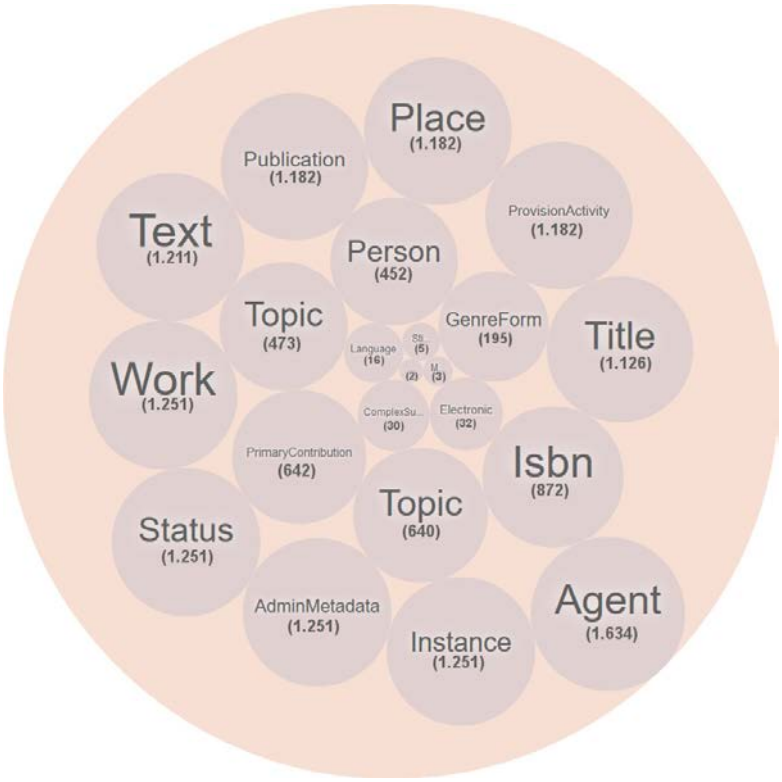
El modelo de conversión que se propone aquí tiene como objetivo la publicación de registros bibliográficos en Linked Data con 5 estrellas. Para ello es necesario que el conjunto de datos resultante esté vinculado con otros datasets con los que tenga relación conceptual.

El principal problema que encontramos en esta fase viene derivado por el software que se emplea. En su versión gratuita el programa GraphDB no permite cargar grandes dumps (ficheros para el volcado de mucha cantidad de información), por lo que no sería posible trabajar, por ejemplo, con VIAF. La prioridad fue la de buscar aquellos datasets que permitieran acceder a sus tripletas a través de un SPARQL Endpoint.

Tabla V. Ejemplo de consulta SPARQL. En concreto se ha empleado para vincular los autores (identificados mediante la clase autorURI) con sus respectivas obras (workURI). Se emplea la url del servicio que asigna automáticamente el software GraphDB (http://localhost:7200/rdf-bridge/1669481892565) para poder lanzar el Sparql Endpoint.

```
PREFIX bf: <http://id.loc.gov/ontologies/bibframe/>
PREFIX bflc: <http://id.loc.gov/ontologies/bflc/>
prefix spif: <http://spinrdf.org/spif#>


INSERT{ GRAPH<http://example.org/>{

   ?workURI a ?tipoR, bf:Work;
      bf:contribution [ a bflc:PrimaryContribution;
      bf:agent ?autorURI] ;

   }
      }
WHERE {
   service <http://localhost:7200/rdf-bridge/1669481892565> {
      ?registroRow a <urn:Row> ;
      <urn:col:001> ?numero ;
      <urn:col:T-Registro> ?tRegistro ;
      <urn:col:Autor-URI> ?autoruri ;


               VALUES (?tRegistro ?tipoR) {
      ("Textual" bf:Text )
      ("Archivo de ordenador" UNDEF )
      ("Material grafico proyectable" bf:MovingImage)
      ("Material grafico bidimensional" bf:StillImage)}

      bind(iri(concat("http://example.org/", spif:encodeURL(?autoruri))) as ?autorURI)
      bind(iri(concat("http://example.org/", ?numero, "#Work")) as ?workURI)
      }
}
```

Figura 2. Clases y subclases obtenidas y número de instancias



Para localizar los datasets que pudieran enriquecer nuestros datos, y que además cumplieran con las condiciones técnicas impuestas por la herramienta escogida, se empleó Datahub (Datahub, 2016). La búsqueda de datasets relacionados con bibliotecas y registros de autoridad, que además estuvieran actualizados y activos, arrojó como resultado:

- Lista de Encabezamientos de materia para las Bibliotecas Públicas (LEMB) – http://id.sgcb.mcu.es/sparql
- Biblioteca Nacional de España (BNE) – http://datos.bne.es/sparql

Como paso previo se realizaron varias consultas desde el SPARQL de GraphDB, sin insertar datos en el repositorio, cruzando los diferentes datos con el objetivo de encontrar significativas y numerosas relaciones entre ambos conjuntos.

De este modo, con la Lista de Encabezamientos de Materia para las Bibliotecas Públicas se localizaron rápidamente vínculos entre la clase Topic de Bibframe y Concept de SKOS, usada por LEMB, concretamente para 211 de los 473 encabezamientos de materias almacenados en el repositorio.

Por su parte la BNE planteó más problemas, y es que en este caso contamos con una colección mucho más amplia en la que se pueden cruzar una mayor cantidad y variedad de datos. Debido al problema anteriormente mencionado del uso de guiones en el código ISBN, no se pudo realizar la conexión con este dataset. A priori esta debía ser la vinculación más atractiva por conectar dos instancias inequívocamente, aunque, por otro lado, no siempre se recuperaban todos los datos de las instancias de la BNE, e incluso, a veces, era imposible recuperar alguno, lo que dificultó todo este proceso.

Por este motivo se recurrió a emplear un dump file de la BNE que establecía una equivalencia entre las URIs de la clase autores de la BNE y su registro en VIAF[1]. Este pudo ser importado a GraphDB, ya que su tamaño no lo impedía y, a través de dos consultas, una para insertarle a cada grafo del dump file el nombre completo del autor y otra para poder buscar entre ellos alguno de los 452 nombres de autores y asociarles su VIAF, se consiguieron encontrar 184 coincidencias.

Una vez localizados y asegurada la existencia de suficientes relaciones, se repitió el proceso visto en la fase de generar: se elaboró un mapa del nuevo

dataset (figura 3), se llevaron a cabo las diferentes consultas SPARQL para insertar dichos datos y se validó con éxito este conjunto de datos.

4.6. Publicación de datos

En este último paso se llevó a cabo la publicación del dataset en la Web. Lo primero que se hizo fue determinar el método de acceso y tecnologías para su publicación. De este modo, se decidió publicarlo a través de una página web, de forma similar a como lo ofrece la Biblioteca Nacional de España, a través de un SPARQL Endpoint y también con un fichero de descarga (dump file).

Para su publicación en un sitio web se pensó en un primer momento en llevarlo a cabo a través de Drupal; no obstante, la falta de módulos, actualización y documentación de los mismos para llevar a cabo tal proceso hizo imposible tomar dicha vía. Se revisaron otros CMS usados por otros proyectos, como Joomla!, pero se obtuvo idéntico resultado.

Es por ello que finalmente se optó por obviar dicha opción y crear en primer lugar el dump file, acompañado de sus metadatos. Es muy importante añadir esto último, con el fin de facilitar la identificación y localización de los datos, conocer su origen y, en definitiva, hacerlos más accesibles (Heath y Bizer, 2011). Lo ideal es crear el conjunto de metadatos lo más exhaustivo posible e insertarlo, por medio de tripletas, en el dataset. Aunque ese fichero descriptivo también debe acompañar al dataset en el sitio web donde se publique.

Por otro lado, se probaron varias de las herramientas más usadas en proyectos Linked Data de bibliotecas, archivos y museos (Smith-Yoshimura, 2016) con la idea de crear un SPARQL Endpoint. Para ello, en un ordenador con Ubuntu 14.04.5 LTS se instaló Openlink Virtuoso 7.2.4, Blazegraph 2.1.4 y GraphDB 8.0.1. El resultado permitió ofrecer una exploración visual del dataset, tal y como se muestra en la figura 4.

5. DISCUSIÓN

Los autores entienden que cuando se propone un modelo de trabajo o una metodología es necesario mostrar su funcionamiento de forma empírica. En este trabajo se demuestra que la propuesta realizada es válida y produce los resultados esperados, pudiéndose aplicar a conjuntos de datos que generan gran cantidad de complicaciones en su gestión, como los encontrados al inicio de nuestra investigación y generando un conjunto de datos en Linked Data con 5 estrellas.

Figura 3. Mapa del vocabulario tras su enriquecimiento con datos de la LEMB y BNE

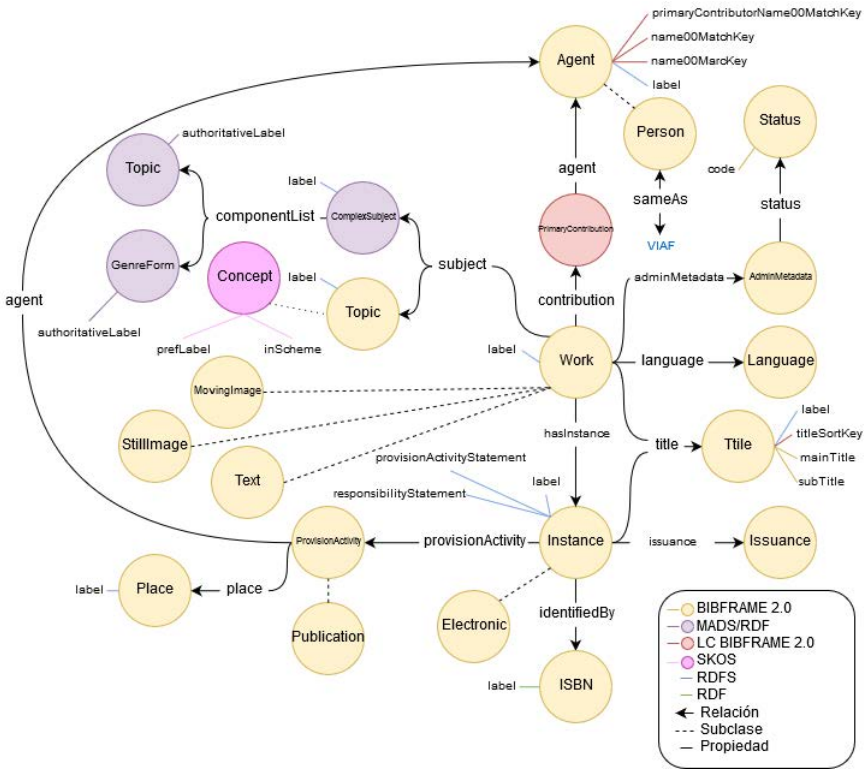
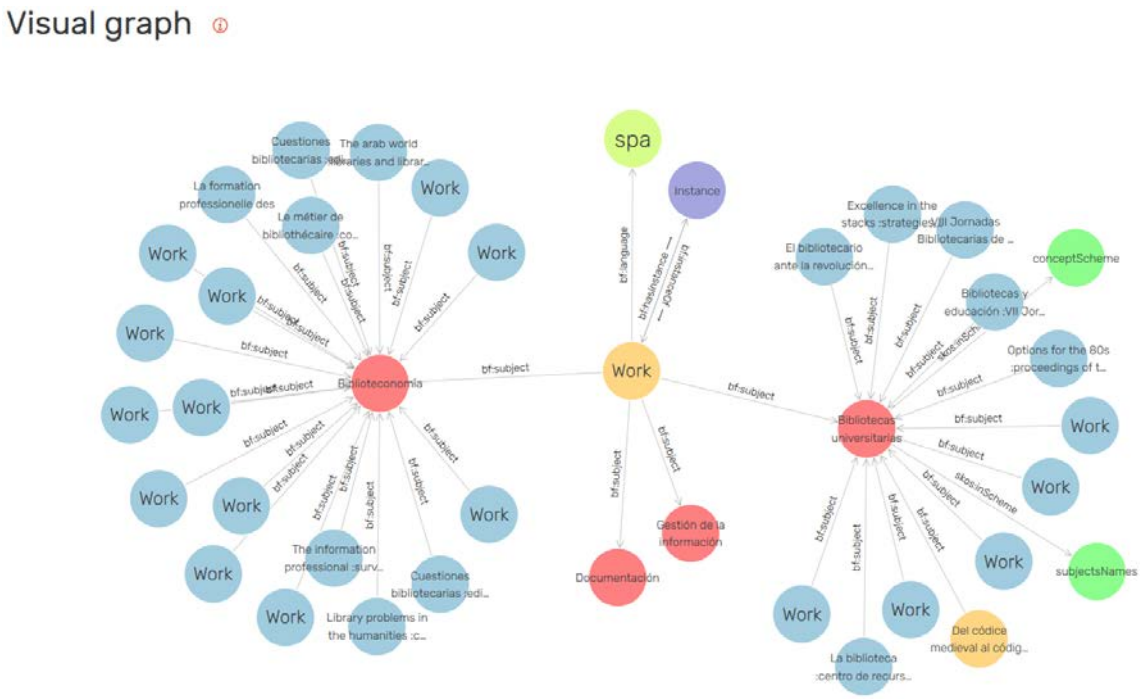


Figura 4. Exploración visual de grafos en GraphDB



Durante la primera fase, cuando se determinó la fuente de datos con la que trabajar, aparecieron varios problemas relacionados con la forma de acceder a los registros. Así, se tuvo que descartar el acceso a los mismos a través del servidor Z39.50 de la BUGR porque este servicio tiene limitada la cantidad de registros a descargar a 500. No logramos entender cuál es la política que ha llevado a esta restricción tan severa, que además no se reproduce en otras bibliotecas universitarias. Pero lo cierto es que este impedimento ha supuesto que se tuvieran que buscar otras vías alternativas que no facilitaban el trabajo, precisamente. Este tipo de servidores son la vía idónea para conversiones de esta clase, ya que ofrecen un archivo ISO 2709 con extensión .mrc (Aenor, 2006), mientas que el OPAC Adrastea genera un fichero en MARC21, con extensión txt.

Una vez realizada la consulta encontramos que, con la misma ecuación de búsqueda, un catálogo (Adrastea) arrojaba 1.251 registros, mientras que otro (Veleta) 1.249. Esto se debió a que en dos registros las materias asociadas no estaban acentuadas, por lo que intuimos que el último buscador parece no aplicar lematizadores, al omitir esos resultados.

A falta de otras vías ofrecidas por la BUGR, la exportación se realizó a través del OPAC Adrastea aunque, de nuevo, encontramos varios problemas. El origen de casi todos se debe a que el fichero MARC generado no es compatible con programas que gestionan este formato. Por lo que se pudo observar a través del editor de texto Notepad++, el fallo se encuentra en la inserción de varios saltos de línea que aparecen al llegar al carácter 40.000, impidiendo con ello una correcta lectura. Para solucionarlo se eliminaron desde este mismo programa dichos saltos, quedando todo el archivo reducido a una única línea.

A pesar de eso, ese fichero en MARC que ya no contaba con problemas en su estructura, ofrecía ahora una no despreciable cantidad de errores que se descubrieron tras llevar a cabo su validación en C# MARC Editor. De hecho, había programas como MarcEdit que directamente no eran capaces de abrir el fichero por la cantidad de errores detectados. El problema lo originó el uso de campos, subcampos e indicadores obsoletos o no existentes en la actualidad, como por ejemplo el desconocido campo 049 o el 019 derivado de IBERMARC para el número de depósito legal. La información que aparecía en muchos de estos campos podía ser obviada, ya que no se encontraba en el lugar correcto, pero afectaba a la correcta gestión del fichero. En total se localizaron errores en 959 registros, un 76% del total.

El resto de formatos de exportación que ofrece el OPAC muestran gran cantidad de inconsistencias que los convierten en inviables, no solo para conversiones de este tipo, sino para cualquier otra cosa. Así, por ejemplo, los registros de pantalla completa que carecen de autor utilizan la mención de responsabilidad del título (245$c) como tal, y

estos junto a los exportados en ProCite incluyen los resúmenes como notas.

El proceso de limpieza de datos permitió localizar muchos problemas derivados, principalmente por una catalogación deficiente. Entre ellos destacan:

- En el campo 008 se encontraron errores en los códigos de idioma, en concreto, aparecieron algunos registros con los términos sp, esp (ambos referidos al español), gao (gallego), ne y ag; teniendo que sustituirlos por aquellos a los que verdaderamente hacían mención.
- En la mayoría de los registros hubo que eliminar caracteres al final y/o principio de la cadena de texto, tales como espacios, puntos o comas, aplicando la misma expresión regular a todos ellos.
- Para el lugar de publicación, en muchas ocasiones, una ciudad podía aparecer con su nombre en varios idiomas. Aunque también se ha dado el caso, muy repetido, de que apareciera el nombre del país al que pertenece una ciudad, pero expresado de diferentes maneras. Si bien es cierto que el campo 260 no requiere de control de autoridades, las especificaciones aportadas por los catalogadores (por ejemplo, indicar la provincia o el estado en el que se encuentra una ciudad) estaban introducidas con errores tipográficos continuos.
- En cuanto a las editoriales, algunas de ellas podían encontrarse con el nombre completo o abreviado a través de sus siglas.
- En las fechas, campo 260$c, el principal problema estaba al eliminar las referencias al depósito legal o copyright en los casos en los que se había extraído de ahí la fecha. Así mismo se encontraron fechas incompletas a las que le faltaba algún dígito.
- El campo 650 fue, sin duda, el que más problemas planteó, ya que su contenido tenía una gran falta de control. Ese descontrol no solo se debía a la falta de normalización de las materias: además se encontraron en ese campo datos que debían corresponder al de notas.

Como se puede observar, las fases de determinación de datos y de limpieza han permitido comprobar el estado de los registros, la calidad de los datos y el nivel de catalogación y normalización aplicado a los mismos. Entendemos que estas fases pueden servir, sin lugar a dudas, como mecanismo para medir la calidad de los registros de cualquier biblioteca, quedando patente en este caso el mal estado de la muestra recogida. Es posible que alguno de los problemas encontrados tenga su origen en un deficiente proceso de conversión del pasado. Cuando la BUGR cambió de programa de automatización se produjo una transformación masiva de registros del antiguo sistema al nuevo. Dicha conversión se llevó a cabo sin demasiado control sobre lo que se generaba, y sin aplicar técnica alguna de limpieza de datos. A esto se le une, como se ha podido atisbar por lo descrito aquí, un trabajo de catalogación bastante deficiente.

Por otro lado, y como se comentó con anterioridad, el hecho de que las bibliotecas no informen sobre el tipo de licencia con la que publican sus datos limita que estos se puedan convertir en vinculables bajo la denominación Linked Open Data. De hecho, el conjunto de registros resultantes del proceso de conversión aquí explicado no se ha podido ofrecer públicamente precisamente por esta limitación. Entendemos que es clave que se ofrezca esta información, ya no solo como elemento que demuestra la propia calidad de los datos, sino también como mecanismo que aumenta la visibilidad de la biblioteca, incrementando las posibilidades de que sus registros sean vinculables.

En lo que respecta a la etapa del modelado, se ha constatado que existe una gran cantidad de vocabularios y ontologías que pueden acomodarse perfectamente a la vinculación de cualquier registro bibliográfico. Ha llamado especialmente la atención el caso del sucesor de MARC, Bibframe, ya que a pesar de estar en constante evolución, cuenta con gran cantidad de documentación que facilita el proceso de convertir registros de un formato a otro. Sin embargo, a la hora de generar los grafos RDF salieron a la luz los principales inconvenientes de este formato. Esos problemas son más patentes a la hora de realizar las consultas en SPARQL, ya que existían campos en muchos registros que no contaban con valores. Aunque la solución, desde el punto de vista técnico, fue sencilla, lo cierto es que no apareció este problema en la extensa documentación consultada.

Para finalizar, entendemos que es necesario que se potencie la creación de herramientas que faciliten el proceso de publicación de los datos. Lo ideal sería que todo el esfuerzo realizado se pudiese ver recompensado con un mecanismo de publicación más dinámico y visual, creando catálogos web que permitan al usuario un uso más atractivo de esa información. Si bien es cierto que la mayoría de proyectos estudiados emplean interfaces web creados ad hoc, en muchos casos esas interfaces cuentan con evidentes problemas desde el punto de vista de la usabilidad web, de tal modo que la información que publican no se encuentra integrada de la mejor forma posible dentro de la web que le da cobijo.

6. CONCLUSIONES

Aunque a estas alturas existen gran cantidad de proyectos de bibliotecas que han convertido sus registros bibliográficos en Linked data con el fin de aprovecharse de las innumerables ventajas que ofrece este sistema de publicación de datos. Lo cierto es que son muchos los factores que han determinado el empleo de diferentes metodologías para lograr dicha conversión. La fuente de datos de partida, los programas empleados, los productos que se desea crear son solo algunos de los muchos condicionantes que han impedido el empleo de un modelo normalizado para realizar esa migración.

Pero, a partir del estudio de todos esos proyectos, en este trabajo se propone una metodología para lograr ese objetivo basada en seis etapas que permite la implementación de esos datos alcanzando las cinco estrellas a las que debería tender cualquier proyecto Linked Data.

Como producto de dicho estudio, y teniendo en cuenta las tareas que están asociadas a estas etapas, también se llega a la conclusión de que existe la posibilidad de automatizar completamente los procesos de extracción de datos, la limpieza y la generación de grafos RDF. Y, lo que es más importante desde el punto de vista de la perdurabilidad de una futura implementación, es que además se puede automatizar este trabajo también con datos nuevos, introducidos en fases posteriores.

Esta metodología ha permitido, por un lado, poner en evidencia la necesidad de que tanto los programas de automatización de bibliotecas como la política bibliotecaria permitan incorporar más métodos para la extracción de datos bibliográficos. Si hace unos años era importante que un software de este tipo ofreciera múltiples opciones de exportación de registros, hoy en día es necesario que se permita el trabajo con un mayor conjunto de datos y que, además, estos cuenten con más calidad. En relación a esto, y por otro lado, la etapa de limpieza de datos -grueso de la implementación y parte fundamental para el éxito del proyecto- además de tener sentido dentro de la metodología propuesta, se ha mostrado como un mecanismo muy válido para verificar y evaluar la calidad de los datos con los que se trabaja y, en concreto, analizar la calidad de las catalogaciones que se almacenan en las bases de datos. En ese sentido, y dentro del caso de estudio al que se ha aplicado la metodología propuesta, se descubre la necesidad de realizar una profunda revisión de los registros bibliográficos, así como de los métodos de acceso y filtrado, especialmente en lo referido al control de autoridades y puntos de acceso, especialmente en la Biblioteca de la Universidad de Granada, pero también en la Biblioteca Nacional de España. De esa manera, se ha puesto de manifiesto el uso incorrecto de códigos en las cabeceras de MARC 21, la descripción de materias y el formato de códigos ISBN en ambas instituciones.

En lo que respecta a la modelación de los datos, es muy posible que muchos de los problemas encontrados tengan solución elaborando un vocabulario u ontología propia. Aunque, si se sigue este camino tal y como se ha hecho en este trabajo con la implementación de varias entidades propias, es necesario realizar una normalización muy clara y documentada. Si, por el contrario, se opta por el empleo exclusivo de Bibframe (algo bastante común en la actualidad) es necesario ser consciente de que se trata de un vocabulario en constante evolución y, aunque cuenta con abundante documentación lo que facilita su uso, lo cierto es que a día de hoy dista mucho de ser un modelo idóneo y que se pueda aplicar de forma genérica a cualquier proyecto.

Para finalizar, es importante destacar la cantidad de herramientas disponibles que permiten acometer prácticamente todas las etapas de esta metodología. Estas aplicaciones son especialmente destacadas en los apartados de conversión de registros bibliográficos, el enriquecimiento y la limpieza. Aunque el principal obstáculo que se encuentra se refiere a la necesidad de disponer de conocimientos en lenguajes de programación, consulta de bases de datos y el trabajo con expresiones regulares. Sin embargo, se echa de menos contar con alguna herramienta que facilite el proceso de publicación a través, por ejemplo, de algún CMS (Content Management System).

Metodologías, como la expuesta en este trabajo, no tendrían sentido si las bibliotecas contaran, a través de su propio programa de automatización, de mecanismos para publicar automáticamente en Linked Data sus registros bibliográficos. De esa manera los datasets ofrecidos estarían al día y no sería necesario realizar un doble esfuerzo, tal y como sucede en la actualidad. Pretender que una biblioteca, con los problemas económicos que arrastra este sector, pueda permitirse el lujo de contar con dos entornos diferentes y gestionados en paralelo carece de sentido. Mientras esto no suceda la mayoría de bibliotecas están abocadas a retrasar su salto hacia Linked Data. Por ese motivo, y en la situación en la que nos encontramos en la actualidad, esta metodología, que es aplicable a cualquier catálogo, tiene razón de ser.

7. NOTAS

1. https://datahub.io/dataset/datos-bne-es/resource/bb29e8ff-5f39-418f-b049-689479ac440a

8. REFERENCIAS

Aenor. (2006). Norma UNE-ISO 2709:2006.

Ávila-García, L.; Ortiz-Repiso, V.; Rodríguez-Mateos, D. (2015). Herramientas de descubrimiento: ¿una ventanilla única? *Revista Española de Documentación Científica*, 38 (1), e077. https://doi.org/10.3989/redc.2015.1.1178

Bermès, E.; Coyle, K.; Dunsire, G. (2011). *Library Linked Data Incubator Group Final Report*. https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/ [consultado el 11-02-2018].

Berners-Lee, T. (2010). Linked Data - Design Issues. https://www.w3.org/DesignIssues/LinkedData.html [consultado el 12/02/2018].

Beastall, G. (2016). The MARC standard format is dying! *Soutron*. https://www.soutron.com/marc-standard-format-bibframe/ [consultado el 07-02-2018].

Biblioteca Nacional de España. (2016a). datos.bne.es. http://datos.bne.es [consultado el 12-02-2018].

Biblioteca Nacional de España. (2016b). Formato *MARC 21 para Registros Bibliográficos*. http://www.bne.es/es/Micrositios/Guias/Marc21/resources/Docs/Marc21.pdf [consultado el 12-02-2018].

Bibliothèque National de France. (2014). data.bnf.fr. http://data.bnf.fr/ [consultado el 12-02-2018].

Cyganiak, R.; Bizer, C. (2011). Pubby – A Linked Data Frontend for SPARQL Endpoints. http://wifo5-03.informatik.uni-mannheim.de/pubby/ [consultado el 12-02-2018].

Cabonero, D.; Dolendo, R. (2013). Cataloging and Classification Skills of Library and Information Science Practitioners in their Workplaces: A Case Analysis. Library Philosophy and Practice. https://digitalcommons.unl.edu/libphilprac/960/ [consultado el 05-02-2018].

Datahub. (2016). DataHub: data online made simple. https://datahub.io/ [consultado el 09-02-2018].

Deliot, C. (2014). Publishing the British National Bibliography as Linked Open Data. *The British Library*. http://www.bl.uk/bibliographic/pdfs/publishing_bnb_as_lod.pdf [consultado el 12-02-2018].

Dimou, A.; Heyvaert, P.; Taelman, R.; Verborgh, R. (2016). Modeling, Generating, and Publishing Knowledge as Linked Data. *Knowledge Engineering and Knowledge Management.* pp. 3-14. Bologna, Italia: Springer.

Europeana. (2017). The European Library Open Dataset. *Europeana.* https://pro.europeana.eu/data/home-data-the-european-library-open-dataset-the-european-library-open-dataset [consultado el 12-02-2018].

Hallo, M.; Luján-Mora, S.; Maté, A.; Trujillo, J. (2016). Current state of Linked Data in digital libraries. *Journal of Information Science*, *42*(2), 117-127. https://doi.org/10.1177/0165551515594729

Hallo, M.; Lujan-Mora, S.; Trujillo, J. (2014). Transforming Library Catalogs into Linked Data. *7th International Conference of Education, Research and Innovation.* pp. 1845-1853; Sevilla, España: IATED.

Harris, S.; Seaborne, A. (2013). SPARQL 1.1 Query Language. https://www.w3.org/TR/sparql11-query/ [consultado el 09-02-2018].

Heath, T.; Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. California: Morgan & Claypool Publishers. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Hidalgo-Delgado, Y.; Senso, J.; Leiva-Mederos, A.; Hípola, P. (2016). Gestión de fondos de archivos con datos enlazados y consultas federadas. *Revista Española de Documentación Científica*, 39(3). https://doi.org/10.3989/redc.2016.3.1299.

Hyvönen, E.; Tuominen, J.; Alonen, M.; Mäkelä, E. (2014). Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. In: Presutti V.; Blomqvist, E.; Troncy, R.; Sack, H.; Papadakis, I.; Tordai, A. (eds.), *The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014. Lecture Notes in Computer Science*, vol. 8798, pp. 226-230. Cham: Springer. https://doi.org/10.1007/978-3-319-11955-7_24

Internet & Data Lab. (2016). Turtle validator. http://ttl.summerofcode.be/ [consultado el 12-02-2018].

Isaac, A.; Waiter, W.; Young, J.; Zeng, M. (2011). Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets. W3C Incubator Group Report 25 october 2011. https://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/ [consultado el 05-02-2018].

Kroeger, A. (2013). The road to Bibframe: the evolution of the idea of bibliographic transition into a post-MARC future. *Cataloguing & Cassification Quartely*, 51 (8), 873-890. https://doi.org/10.1080/01639374.2013.823584

Library of Congress. (2012). *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services.* Washington. https://www.loc.gov/bibframe/pdf/marcld-report-11-21-2012.pdf [consultado el 08-02-2018].

Library of Congress. (2015). MADS/RDF Documentation. http://www.loc.gov/standards/mads/rdf/ [consultado el 12-02-2018].

Library of Congress. (2016a). LC BIBFRAME 2.0 Vocabulary Extension List View. http://id.loc.gov/ontologies/bflc.html [consultado el 12-02-2018].

Library of Congress. (2016b). LC Linked Data Service: Authorities and Vocabularies (Library of Congress). http://id.loc.gov/ [consultado el 12-02-2018].

Library of Congress. (2017a). MARC 21 to BIBFRAME 2.0 Conversion Specifications (BIBFRAME - Bibliographic Framework Initiative, Library of Congress). https://www.loc.gov/bibframe/mtbf/ [consultado el 12-02-2018].

Library of Congress. (2017b). BIBFRAME - Bibliographic Framework Initiative. https://www.loc.gov/bibframe/ [consultado el 12-02-2018].

LOV. Linked Open Vocabularies. (2016). Linked Open Vocabularies. http://lov.okfn.org/dataset/lov/ [consultado el 27-09-2017].

MacKenzie, S.; Carl, G.; Stahmer, X. L.; Gloria, G. (2017). *BIBFLOW: A Roadmap for Library Linked Data Transition*. https://bibflow.library.ucdavis.edu/wp-content/uploads/2017/03/bibflow_roadmap_revised_3_14_2017.pdf [consultado el 09-02-2018].

Marais, H. (2009). *Authority control in an academic library consortium using a union catalogue maintained by a central office for authority control*. Tesis doctoral. Pretoria: University of South Africa. http://hdl.handle.net/10500/2546 [consultado el 05-02-2018].

Marcum, D. (2011). A bibliographic framework for digital age. *Library of Congress*. https://www.loc.gov/bibframe/news/framework-103111.html

Montalvillo Mendizabal, L. (2012). *Definición y desarrollo de herramienta web de gestión de metadatos Business Intelligente* [tesis de maestría]. Barcelona: Universidad Politécnica de Cataluña. https://upcommons.upc.edu/handle/2099.1/16145 [consultado el 09-02-18].

Morris, T. (2015). General Refine Expression Language. https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language [consultado el 05-02-2018].

Ontology Engineering Group. (2017). vocab.linkeddata.es. http://vocab.linkeddata.es/ [consultado el 10-02-2018].

OpenLink. (2015). Virtuoso Linked Data. https://virtuoso.openlinksw.com/linked-data/ [consultado el 12-02-2018].

Papadakis, I.; Kyprianos, K.; Stefanidakis, M. (2015). Linked Data URIs and Libraries: The Story So Far. *D-Lib Magazine*, *21*(5/6). https://doi.org/10.1045/may2015-papadakis

Peset, F.; Ferrer-Sapena, A.; Subirats-Coll, I. (2011). Open data y Linked open data: su impacto en el área de bibliotecas y documentación. *El Profesional de la Información*, 20 (2), 165-173. https://doi.org/10.3145/epi.2011.mar.06

Smith-Yoshimura, K. (2016). Analysis of International Linked Data Survey for Implementers. *D-Lib Magazine*, 22 (7/8). https://doi.org/10.1045/july2016-smith-yoshimura

Snow, K. (2011). *A Study Of The Perception Of Cataloging Quality Among Catalogers In Academic Libraries*. Tesis doctoral. University of North Texas. http://digital.library.unt.edu/ark:/67531/metadc103394/ [consultado el 05-02-2018].

Stumpf, F. F. (2003) Centralized cataloging and processing for public library consortia. *The Bottom Line*. 16(3). https://doi.org/10.1108/08880450310488003 [consultado el 05-02-2018].

Subirats I.; Malapela, T.; Dister, S.; Zeng, M.; Goovaerts, M.; Pesce, V.; Jaques, Y.; Anibaldi, S.; Keizer, J. (2012). Reorienting Open Repositories to the Challenges of the Semantic Web: Experiences from FAO's Contribution to the Resource Processing and Discovery Cycle in Repositories in the Agricultural Domain. En: Dodero, J.M.; Palomo-Duarte, M.; Karampiperis, P. (eds.), *Metadata and Semantics Research*. Springer: Berlin, Heidelberg. Vol 343, 158-167. https://doi.org/10.1007/978-3-642-35233-1_17

Sulé, A.; Centelles, M.; Franganillo, J.; Gascón, J. (2016). Aplicación del modelo de datos RDF en las colecciones digitales de bibliotecas, archivos y museos de España. *Revista Española de Documentación Científica*, 39(1), e121. https://doi.org/10.3989/redc.2016.1.1268

Taylor, S.; Jekjantuk, N.; Mellish, C.; Pan, J. Z. (2013). Reasoning Driven Configuration of Linked Data Content Management Systems. *Joint International Semantic Technology Conference - JIST 2013.* pp. 429-444: Springer, Seoul.

Tennant, R. (2002). MARC must die. *Library Journal,* 127 (17), 26-28. http://soiscompsfall2007.pbworks.com/f/marc+must+die.pdf [consultado el 07-02-2018].

The British Library. (2014). Welcome to bnb.data.bl.uk. http://bnb.data.bl.uk/ [consultado el 12-02-2018].

Torre-Bastida, A.-I.; González-Rodríguez, M.; Villar-Rodríguez, E. (2015). Datos abiertos enlazados (LOD) y su implantación en bibliotecas: iniciativas y tecnologías. *El Profesional de la Información*, *24*(2), 113-120. https://doi.org/10.3145/epi.2015.mar.04

Vila-Suero, D.; Gómez-Pérez, A. (2013). datos.bne.es and MARiMbA: an insight into Library Linked Data. *Library Hi Tech*, 31(4), 575-601.

Vila-Suero, D.; Villazon-Terrazas, B.; Gomez-Perez, A. (2012). datos.bne.es: A library linked dataset. *Semantic Web*, 4, 307-313.

Volz, J.; Bizer, C.; Gaedke, M.; Kobilarov, G. (2009). Silk – A Link Discovery Framework for the Web of Data. *2nd Workshop about Linked Data on the Web (LDOW2009)*, Madrid. http://events.linkeddata.org/ldow2009/papers/ldow2009_paper13.pdf [consultado el 10-02-2018].

Wang, Y.; Stash, N.; Aroyo, L.; Gorgles, P.; Rutledge, Ll.; Schreiber, G. (2008). Recommendations based on semantically enriched museum collections. *Web Semantica: Science, Services and Agents on the World Wide Web,* 6(4). https://doi.org/10.1016/j.websem.2008.09.002 [consultado el 06-02-2018].

Wenz, R. (2013). Linked open data for new library services: the example of data.bnf.fr. *Italian Journal of Library, Archives and Information Science*, *4*(1). https://doi.org/10.4403/jlis.it-5509

Wolverton, R. E. (2005). Authority Control in Academic Libraries in the United States: A Survey. Cataloging & Classification Quarterly, 41(1), 111-131. https://doi.org/10.1300/J104v41n01_06