Revista Española de Documentación Científica
39(3), julio-septiembre 2016, e145
ISSN-L:0210-0614. doi: http://dx.doi.org/10.3989/redc.2016.3.1299

ESTUDIOS / *RESEARCH STUDIES*

Gestión de fondos de archivos con datos enlazados y consultas federadas

Yusniel Hidalgo-Delgado*, José A. Senso**, Amed Leiva-Mederos***, Pedro Hípola**

* Departamento de Programación, Universidad de las Ciencias Informáticas, La Habana, Cuba
** Departamento de Información y Comunicación, Universidad de Granada, Granada
*** Universidad Central «Marta Abreu» de Las Villas, Cuba
Correos-e: yhidalgo@uci.cu; jsenso@ugr.es; amed@uclv.edu.cu; phipola@ugr.es

Recibido: 24-04-2015; 2ª versión: 11-09-2015; Aceptado: 26-10-2015

Cómo citar este artículo/Citation: Hidalgo-Delgado, Y.; Senso, J. A.; Leiva-Mederos, A.; Hípola, P. (2016). Gestión de fondos de archivos con datos enlazados y consultas federadas. *Revista Española de Documentación Científica*, 39(3): e145. doi: http://dx.doi.org/10.3989/redc.2016.3.1299

Resumen: En este trabajo se presentan las principales tecnologías de la Web Semántica que pueden ser de utilidad para la gestión de fondos archivísticos. Se examinan diversos proyectos de ámbito internacional y local que parten de descripciones normalizadas ISAD-G para generar ontologías, así como la disponibilidad de LIAM (Linked Archival Metadata), que facilita la transformación de datos de archivo a formato RDF (Resource Description Framework). Por otra parte, se analiza cómo la gestión de datos enlazados permite la interoperabilidad entre sistemas de información y la búsqueda facetada a partir de fondos documentales almacenados, descritos en OWL (Ontology Web Language), SKOS (Simple Knowledge Organization System) y Dublin Core. Los autores proponen la utilización de un CMS (Content Management System) que gestione fondos de archivo, compatible con SIOC (Semantically-Interlinked Online Communities) y OAI-PMH (Open Archives Initiative - Protocol Metadata Harvesting), para facilitar el intercambio y la recuperación de información. En concreto, se detallan las tecnologías que se han utilizado para desarrollar CoroArchivo, sistema que además se evalúa con un experimento que realiza la creación automática de ontologías a partir de descripciones ISAD-G almacenadas en DSpace. La herramienta desarrollada permite realizar consultas federadas sustentadas en las clases de exclusión e igualdad del vocabulario OWL.

Palabras clave: Datos enlazados; ontologías; archivos; repositorios; servicios de información; Drupal; búsqueda federada; DSpace.

Management of archival materials with Linked Data and federated queries

Abstract: In this paper the major technologies of the Semantic Web which may be useful for archives management are summarized. Several local and international projects that generate ontologies from standardized descriptions based on ISAD-G are examined. It is also discussed LIAM (Linked Archival Metadata), that facilitates the transformation of archive records into RFD (Resource Description Framework) format. Furthermore, we analyze how Linked Data enables interoperability between information systems and faceted search of OWL (Ontology Web Language), SKOS (Simple Knowledge Organization System) and Dublin Core records. The authors propose the use of a CMS (Content Management System) compatible with SIOC (Semantically-Interlinked Online Communities) and OAI-PMH (Open Archives Initiative - Protocol for Metadata Harvesting) for archive records to improve the exchange and retrieval of information. We specifically describe the technologies used for developing CoroArchivo, system assessed by an experiment that automatically generates ontologies from ISAD-G records stored in DSpace. The evaluation tool lets users perform federated queries based on the OWL vocabulary disjointness and equivalent classes.

Keywords: Linked Data; ontologies; archives; repositories; information services; Drupal; federated queries; DSpace.

1. INTRODUCCIÓN

La Web Semántica (Berners-Lee y otros, 2001) se ha ido implantando en diferentes ámbitos de la actividad documental, como el procesamiento automatizado de la información, el control de autoridades, la recuperación de datos enlazados, los servicios de resumen... Y parece claro que diversas iniciativas de la Web Semántica pueden ser de gran utilidad en la gestión archivística (Mena, 2006; Hernández, 2007). Una de las iniciativas que más recorrido tiene dentro del ámbito de la gestión de información es Linked Data, o datos enlazados, entendido como el conjunto de buenas prácticas para la publicación de datos con el fin de facilitar su reutilización. Esta idea poco a poco se va asentando en la comunidad archivística.

Tanto es así que se han realizado aproximaciones desde distintos puntos de partida. Sin pretender realizar aquí un recorrido exhaustivo, ya que no es el objetivo principal de este trabajo, podemos agrupar en varias corrientes las diferentes formas de afrontar la publicación de datos enlazados en el terreno de los archivos. Por un lado, aquellos proyectos que intentan introducir elementos semánticos en los registros para su posterior conversión a RDF. Por otro, aquellos que se centran principalmente en el control de autoridades y materias mediante diversos mecanismos de transformación.

Dentro del primer grupo, uno de los trabajos más destacados es el de Vasallo que, de forma teórica, muestra cómo realizar la transformación de la norma Topic Maps en RDF (Vasallo, 2010). Este proceso, dentro del ámbito archivístico, transforma elementos de la norma ISAAR (CPF) en relaciones Scope y asociaciones Topic Map. Aunque se trata de un caso muy puntual y difícilmente extrapolable a otras realidades, es interesante porque pone de manifiesto algunos puntos de contacto entre la semántica y la descripción archivística. Evidentemente sus limitaciones en el ámbito de la gestión de autoridades y de entidades externas hacen que sea difícil desarrollar la propuesta. Gracy, por su parte, plantea una metodología que permita la transformación del formato EAD en Linked Data, estableciendo un modelo de datos para que el formato archivístico pueda interoperar con clases y propiedades de ontologías específicas (Gracy, 2014). La autora propone alinear etiquetas EAD con la DBpedia, detectando aquéllas que puedan ser utilizadas en los procesos de descripción archivística, con el fin de enriquecer los registros. En una segunda fase se establecen coincidencias con los vocabularios FOAF, el formato LOD (Linking Open Description of Events) y un conjunto de metadatos geográficos. Los resultados de esta metodología muestran las enormes posibilidades que ofrece EAD en combinación con otros vocabularios ontológicos y su salida Linked Data. Para gestionar el vocabulario se emplea la herramienta de procesamiento de lenguaje natural Open Calais, actualmente propiedad de Thomson Reuters. El principal problema de la arquitectura propuesta por Gracy reside en que genera constantes inconsistencias en los registros, a causa de sus escasos mecanismos de limpieza de datos, con lo que se complica mucho el proceso de reutilización posterior. Además, no se ha verificado la efectividad del proceso por medio de ningún mecanismo de evaluación. Algo similar es lo que sucede con el trabajo de Nam-Park, que transforma los registros del Archivo Nacional de Korea en datos enlazados usando SKOS y Dublin Core para facilitar la conversión en RDF, lo que favorece la realización de búsquedas flexibles (Nam-Park, 2015). El gran logro de este proyecto, desde nuestro punto de vista, reside en lograr la interoperabilidad entre aplicaciones al ofrecer servicios de información con datos extraídos tanto de una biblioteca, un museo y los archivos, a través de la conexión de varias bases de datos. Sin embargo, el resultado genera registros que no pueden ser enlazados, bien por estar incompletos, bien por falta de enlaces internos.

Consideramos que el trabajo más valioso en materia de transformación de datos es el de Rademaker, que propone una arquitectura funcional para realizar la transformación a Linked Data empleando para ello los registros de la Fundación Getulio Vargas (Rademaker y otros, 2015). La clave del proceso está en la migración de datos, que comienza con la transformación de una base de datos relacional a OAI-PMH, permitiendo la exportación a RDF mediante la herramienta D2RQ. Esto facilita el mapeo de los datos y la construcción de un grafo que, después de alinearlo con varios vocabularios (FOAF, SKOS, Dublin Core y Prov), se almacena en una tripleta y es indexado con Apache Solr (servidor open source en Java para realizar búsquedas). Los datos pueden ser consultados mediante SPARQL.

La segunda aproximación se centra en proponer un cambio estructural para la sintaxis del registro de archivo. En ese sentido el trabajo que en 2010 realizó el EACWG (Encoded Archival Context Working Group) con EAC-CPF (Encoded Archival Context – Corporate Bodies, Persons, and Families) marca un punto de inflexión, al corregir los errores relacionados con el control de autoridades dentro de EAD. Esta nueva norma facilita la interoperabilidad al proporcionar un mecanismo que permite desarrollar la norma ISAAR (CPF) en formato de ontología. Además de controlar autoridades, esta pauta permite generar enlaces a recursos de autoridad externos, empleando para ello un subconjunto de etiquetas de la recomendación Xlink.

Muy relacionado con EAC-CPF se encuentra el proyecto SNAC (Social Networks and Archival Context), desarrollado por Daniel Pitti, que perfecciona los mecanismos para el control de autoridades y gestión de puntos de acceso (Lynch, 2014). Para ello extrae las etiquetas EAC-CPF de los registros EAD, y se mapean con diferentes esquemas (VIAF, Library of Congress Name Authority File y Union List of Artist Names). De esta manera las entradas de autoridad no solo son controladas por EAC-CPF, sino que existirán entradas alternativas para los títulos a través de VIAF (Virtual International Authority File), y para datos históricos mediante los otros dos sistemas. SNAC crea además un prototipo para acceder a recursos de archivo con carácter histórico a partir de los datos de redes sociales y profesionales, enlazando recursos de archivos, bibliotecas y museos.

Más allá de la simple conversión de datos en RDF, han aparecido estudios que analizan la gestión de vocabularios controlados para buscar la interoperabilidad y mejorar la búsqueda y la recuperación de la información. En este sentido destaca el trabajo de Grimoüard, que estudia las posibilidades para la gestión de materias mediante endpoints de SPARQL en archivos franceses (Grimoüard, 2014).

Hay que destacar, para finalizar, el desarrollo de LIAM (Linked Archival Metadata), un formato de metadatos para la transformación de descripciones archivísticas en RDF, que facilita las trasformaciones que se hacían sobre EAD para la conversión de datos de archivo, así como los esfuerzos de EGAD (Expert Group on Archival Description) por lograr una nueva versión de la norma ISAD-G más relacional.

La Web Semántica enriquece a la archivística no solo aumentando el valor de las descripciones, como se refleja en la mayoría de trabajos aquí descritos, o en proyectos como Europeana o Archive Hub. Estas iniciativas dibujan también una realidad: los datos de los archivos cobran más relevancia que un simple registro descriptivo, puesto que se aporta la posibilidad de la reutilización, revalorizando el trabajo que los profesionales realizan.

La principal ventaja de LOD está en la reutilización, por lo que la forma en la que se consiga llegar a ella tampoco ha de ser tan determinante, aunque queda claro que este proceso debería evitar pérdidas de información y mantener, en la medida de lo posible, la descripción realizada y las relaciones del objeto descrito con su entorno. Y de ahí parte nuestro trabajo: sobre la base de lo explicado anteriormente, hemos desarrollado una aplicación, CoroArchivo, que facilita la descripción de documentos de archivo en universidades y la gestión de la información mediante Linked Data. El sistema emplea Drupal como CMS, y DSpace como proveedor de datos. Se trata de una herramienta de trabajo en entorno web capaz de gestionar grandes conjuntos de datos transformados en descripciones RDF, y que tiene como principal objetivo el procesamiento y el intercambio de información entre archivos de universidades.

2. TECNOLOGÍAS DISPONIBLES

2.1. OAI-PMH

Existen diferentes modelos para el tratamiento de datos enlazados dentro de repositorios OAI-PMH. En la mayoría de los casos las plataformas empleadas, el software soportado o los proyectos de aplicación, suelen determinar cómo proceder con cada uno de ellos. Resultaría realmente complejo categorizar todas las propuestas que existen, ya que cada una de ellas se adapta a la realidad tecnológica con la que tiene que trabajar, así como a la naturaleza de los datos. Pero sí destacaremos alguno que nos ha servido a modo de guía o patrón, como por ejemplo el trabajo de Coppens, que aporta una forma muy novedosa de gestionar fondos de archivo empleando datos enlazados por medio de los estándares de descripción EAD e ISAD-G (Coppens y otros, 2009). En el mismo sentido, la forma en la que Europeana ha integrado los materiales de archivos, bibliotecas y museos (Haslhofer y otros, 2011; Doerr y otros, 2011), o Nestor, una aplicación construida a partir de la experiencia práctica de varios archiveros, y que emplea EAD para los metadatos (Ferro y Silvello, 2013), son ejemplos clave.

Existe un grupo de iniciativas que, para preservar datos digitales, ha optado por OAI-PMH por las ventajas que presenta este modelo. Nos hemos centrado en el estudio de Preso, que busca la creación de espacios para preservar materiales fílmicos y fotográficos. Por su ámbito de actuación, la Unión Europea, se orienta a ofrecer soporte para Europeana (Addis y otros, 2010).

Aunque el uso de OAI-PMH está bastante generalizado, también se han estudiado aquellos proyectos que proponen complementar este protocolo por medio de diferentes mecanismos. Así OAI2LOD es un servidor, compatible con protocolos de interoperabilidad, que permite consultar los metadatos con SPARQL (Sparql Protocol and RDF Query Language). Realmente esto supone un salto cualitativo con respecto al resto de proyectos, ya que resulta posible sacar más partido a las relaciones existentes entre los datos. En concreto el trabajo con las relaciones "same as" entre dos instancias

del servidor aporta una mayor riqueza en las búsquedas (Haslhofer y Schnadl, 2008, 2010). Para lograr esto, el software servidor compara los valores de un conjunto de atributos seleccionados manualmente según su similaridad léxica usando la distancia Levensthein. Es un sistema modular, ya que para la transformación de los datos en RDF/XML se emplea otro servidor que trabaja de forma paralela, D2RQ Server (Bizer y Seaborne, 2004). Uno de los principales problemas detectados en OAI2LOD, la imposibilidad de compaginar varios proveedores de datos, se resolvió con importantes mejoras realizadas por Coppens, que permitieron la importación de datos procedentes de diversos repositorios OAI-PMH (Coppens y otros, 2013).

2.2. Búsqueda federada

Una de las áreas más investigadas dentro del ámbito de la Web Semántica ha sido la recuperación de información. Entre las primeras aproximaciones destaca Sindice, que embebe datos RDF y los microformatos que los describen en un servidor que se consulta por medio de un SPARQL Endpoint, permitiendo realizar consultas semánticas sobre los recursos almacenados (Tummarello y otros, 2007). Otras aportaciones que han servido como base para el trabajo que se describe en este texto es el método de construcción de índices en tiempo real que emplea Squin (Hartig y otros, 2009), y las consultas basadas en reglas que emplea FedX, que por medio del framework Sesame permite la optimización de consultas federadas (Schwarte y otros, 2011).

En esa misma línea, pero empleando técnicas diferentes, se encuentra el motor Darq, que complementa ese tipo de búsquedas con un sistema de toma de decisiones para refinar las búsquedas federadas (Quilitz y Leser, 2008). Existen otros sistemas que se valen de información estadística para realizar procesos similares, como Splendid (Görlitz y Staab, 2011). El problema es que ambos métodos se centran demasiado en técnicas complementarias y no exploran lo suficiente las facilidades semánticas de recuperación de información a través del lenguaje OWL. Precisamente por ese motivo consideramos tan destacada la aportación de Coppens, que gestiona datos enlazados mediante un índice estructurado para optimizar las consultas *OWL:same as* en los Endpoints locales y remotos de SPARQL de forma simultánea (Coppens y otros, 2013).

También ha servido de base para el proyecto CoroArchivo el motor de consultas Elite, que permite mapear con gran facilitad las relaciones entre los datos por medio de estructuras R-Tree, integrando información de diversa índole (incluidas tripletas RDF) con bases de datos relacionales (Nolle y Nemirovski, 2013).

La combinación de estas tecnologías ha servido para desarrollar el sistema de búsqueda federada de nuestro proyecto, enriqueciendo las consultas del tipo *OWL:same as*. Se ha optado por mapear las clases que se obtienen de la búsqueda distribuida para filtrar mejor la información. La similaridad de los datos se obtiene mediante el cálculo de la distancia Manhattan. Los resultados se complementan con las etiquetas *OWL:equivalentClass*, *OWL:equivalentProperty*, *OWL:differentFrom* y *OWL:AllDifferent*, combinadas con el algoritmo de OWL2tips (Domínguez-Velasco, 2013a). Al igual que en Elite, nuestra propuesta emplea un razonador para formular las consultas y un módulo de indexación distribuida que conecta el sistema con ontologías, tanto locales como externas, a partir de mapas de referencias.

3. COMPONENTES DEL SISTEMA

3.1. Herramientas DSpace para el trabajo con Linked Data

Hemos utilizado el software open source DSpace, que gestiona repositorios compatibles OAI-PMH usando un solo data center.

A pesar de las deficiencias que se le atribuyen a la filosofía OAI-PMH, porque, al solo contar con el protocolo de metadatos OAI-DC (Schöpfel y otros, 2012), gestiona con austeridad las descripciones bibliográficas, y porque a veces forma parte de la web invisible y resulta poco accesible para los servicios de búsqueda (Merlino-Santesteban, 2012), DSpace ha ganado popularidad cuando gestiona datos semánticos sobre todo gracias a:

- el uso de gramáticas de contexto entre repositorios
- la gestión y distribución de todo tipo de metadatos
- la posibilidad de que se recupere información de varios repositorios con una misma interfaz
- su capacidad para generar búsquedas estructuradas y consultas remotas de datos enlazados
- sus funcionalidades para publicar datos enlazados
- las posibilidades que ofrece para construir léxicos, tesauros y ontologías.

3.2. Diseño del sistema

3.2.1. Transformación y desarrollo de ontologías para DSpace

Hemos partido de tres ontologías:

- una en SKOS (Solomou y Papatheodorou, 2010) con funciones de tesauro
- otra para FOAF (Brickley y Miller, 2010), que sirve de herramienta de control de autoridades
- una tercera especial para la norma ISAD-G en OWL, que enriquece la descripción archivística y aprovecha las propiedades de la estructura semántica en la búsqueda federada.

A continuación se explica el proceso de generación de una ontología en ISAD-G y sus vocabularios.

La descripción archivística ISAD-G gira en torno a normas de procesamiento documental que podríamos considerar algo alejadas de otros formatos que se utilizan habitualmente en la Web Semántica. Para construir una ontología que procesara documentos de archivo, partiendo de registros ISAD-G, se conservaron todos los elementos de la descripción transformándolos a formato OWL. Para ello se siguió el mismo proceso planteado en anteriores proyectos (Koutsomitropoulos y otros, 2008; Baker, 2012). Se utilizaron las propiedades de los objetos y de los datos junto a sus especificaciones de cardinalidad. La ontología generada en OWL fue convertida a XML, facilitando la transformación de los elementos de la descripción archivística en el entorno del repositorio DSpace, como en otros proyectos similares (Koutsomitropoulos y otros, 2008). Lo primero que se realiza en este tipo de conversión es la transformación de la sintaxis. A continuación la semántica.

- Transformación de la sintaxis: se partió de la estructura sintáctica de OWL, de SKOS y de Dublin Core para ajustar los datos a la estructura de DSpace. En esta etapa es necesario mostrar todos los datos de OWL en XML, sin que se pierda la riqueza de las descripciones al integrarlas en Protégé, el software gestor de ontologías elegido en nuestro caso.

  Lo primero que se hizo fue construir tablas de convergencia de datos para que las clases principales de los vocabularios no se perdieran. Se decidió utilizar Sampras (Domínguez-Velasco, 2013b), sistema capaz de conservar las propiedades de todos los "datatypes" usados en el diseño del modelo ontológico, estructurar los "syntax encoding schemes" y "cosificar" los datos de Protégé. Con SKOS se mejoró el control del vocabulario; de FOAF para propiciar el control de autoridades. OWL sirvió para buscar relaciones de inclusión y exclusión a través de elementos útiles en la recuperación de la información. El proceso permitió asignar los valores exactos de cada campo en el modelo de OAI-PMH, lo cual se logra con indicadores de relación, reglas de construcción o constructores y refinamiento de datos. Concluido el proceso de cosificación de datos, se importó el resultado dentro de DSpace, que genera un archivo para su visualización en formato XML donde subyacen las estructuras de los tres lenguajes (tabla I).

Tabla I. Equivalencia de la ontología en SKOS

Elemento de XML	Función	Notación de SKOS
TC	Describe un término descrito baja entrada principal	<skos:Concept>
ET	Traducción del Término al inglés	<skos:prefLabel lang="en">
RT	Traducción del Término al Ruso	<skos:prefLabel lang="ru">
PT	Traducción del Término al portugués	<skos:prefLabel lang="po">
ET'	Traducción alternativa al inglés	<skos:altLabel lang="en">
RT'	Traducción alternativa al ruso	<skos:altLabel lang="ru">
PT'	Traducción alternativa al portugués	<skos:altLabel lang="po">
TR	Término relacionado	<skos:related>
NA	Pequeña descripción	<skos:definition>
DC	Clasificación Dewey	<skos:notation>
USE	Remite de un término no autorizado a uno autorizado	<skos:altLabel lang="el">

- Construcción semántica: para obtener una ontología acorde a las especificaciones semánticas de DSpace y establecer constructores y restricciones que evitaran solapamientos, se modelaron las jerarquías dentro de cada vocabulario, de forma que no se solaparan las clases y las subclases con las instancias. Se identificaron las dimensiones de cada ítem, de modo que, una vez cargado el vocabulario, no se generara ruido al gestionar los conceptos de clases y anotaciones con otros elementos organizativos de DSpace. Se expresaron correctamente las relaciones de similaridad léxica entre elementos no registrados en los vocabularios RDF para que no se mezclaran, dentro de un mismo léxico, con otros del sistema o con la estructura de DSpace. Finalmente, fue imprescindible modelar las relaciones de asociación empleando restricciones y operadores de roles. Este paso conllevó la estructuración de los contenidos usando la propuesta de estructura RDF de LIAM desarrollada específicamente para fondos de archivo dentro de un repositorio.

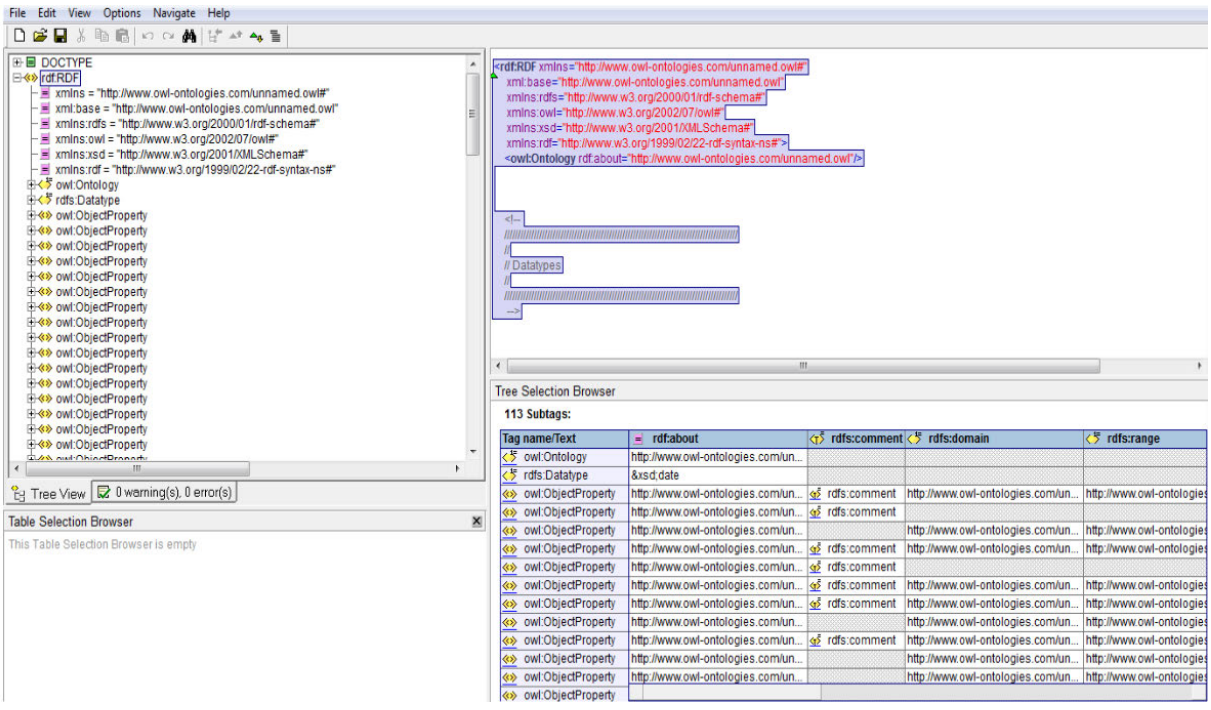
La ontología, una vez adecuada sintáctica y semánticamente, permite que sus datos OWL se exporten con DSpace a XML. El resultado obtenido se transforma en nodos de esquema de DSpace (figura 1).

3.2.2. Uso de datos enlazados a partir de las tecnologías OAI-PMH

Las tecnologías de la Web Semántica han evolucionado hasta generar plataformas de elevado nivel para el tratamiento de datos RDF. La plataforma CoroArchivo incorpora varias de las condiciones necesarias para la publicación de datos enlazados, así como facilidades para la conversión de estos datos, entre las que se encuentran: extracción de metadatos, procesamiento, modelado de datos, transformación, enlace, publicación y normalización. Las herramientas utilizadas han sido:

- Authoris (Leiva-Mederos y otros, 2013) para transformar datos de ontologías en registros de autoridad y gestionar datos enlazados utilizando las prestaciones de Silk, también en "triple stores" descargados de formatos HTML.
- Pubby (Cyganiak y Bizer, 2008), que facilita la conexión y gestión de los triple stores conectándolos con fuentes de datos externas, sin que resulte necesario lanzar constantes consultas SPARQL a diversos Endpoints. También facilita la visualización de datos en HTML y las consultas de tripletas en este formato. Esta aplicación trabaja en Java, por lo que es necesario emplear el software servidor Apache Tomcat.

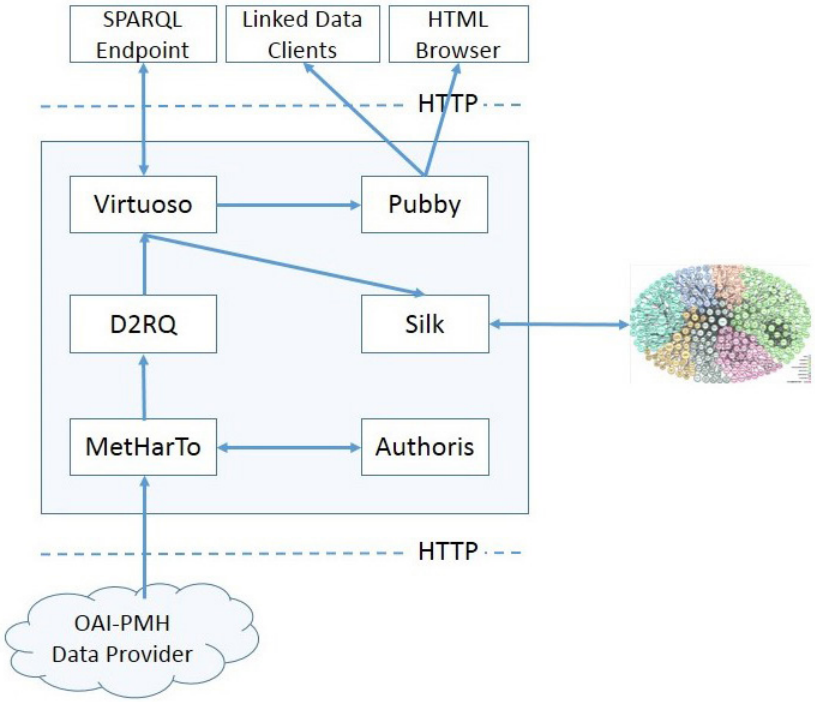
Figura 1. Estructura xml de la ontología en ISAD-G (Ledesma, 2013)



- Silk (Jentzsch y otros, 2010) para la construcción de enlaces de datos en recursos disímiles, uniendo los enlaces RDF de las fuentes particulares con otras fuentes de datos. Dispone de una consola para realizar enlaces entre dos datasets y un servidor http capaz de recibir datos e introducirlos en el flujo RDF a través de "data ítems". La flexibilidad y calidad de los datos que gestiona el lenguaje Silk-LSL permite adaptarlos a la filosofía de trabajo Linked Data.
- Metharto (Hidalgo-Delgado y otros, 2013), capaz de extraer los metadatos de archivo y los elementos de SKOS e ISAD-G del proveedor de datos soportado en OAI-PMH.
- OpenLink Virtuoso (Haslhofer y Schnadl, 2008), servidor universal que combina las funcionalidades de los tradicionales gestores de bases de datos relacionales, bases de datos virtuales, RDF, XML, y que facilita el uso de texto libre en aplicaciones web.
- D2RQ (Bizer y Seaborne, 2004), lenguaje de mapeo que lleva a cabo la relación entre bases de datos y lenguajes semánticos RDF y vocabularios OWL, capaz de transformar un documento RDF en la estructura sintáctica de turtle. Los mapas que se generan definen un grafo en RDF donde se incluye toda la información de la base de datos al igual que en SQL, solo que los datos generados en la estructura del RDF siempre van a exigir tablas relacionales virtuales. Se puede acceder a la plataforma RDF de varias formas según las necesidades de implementación. Por ejemplo: acceso a SPARQL, al servidor para datos enlazados, a un almacén de datos RDF, a una interfaz simple para RDF que convierte los datos bibliográficos a formato LIAM, y a una API de Jena que interopera bases de datos DR2Q.

El flujo de trabajo en el tratamiento de las ontologías y la implementación de Linked Data ha sido el siguiente. Metharto extrae del proveedor de datos las descripciones bibliográficas necesarias y se comunica con Authoris para determinar las entradas de autoridad a partir de reglas de decisión (Leiva-Mederos y otros, 2013). D2RQ genera el mapa del archivo con todos los campos y tablas organizados en un esquema relacional. Las tablas son la equivalencia de las clases, subclases e instancias previamente diseñadas en las ontologías y exportadas a DSpace. Con el script para RDF denominado Dump-rdf se genera un grafo que se aloja en Virtuoso, servidor que provee triple stores y permite consultas SPARQL. Teniendo los datos en Virtuoso se puede configurar Silk para obtener enlaces en lenguaje OWL a través de todos los grafos del sistema, de VIAF, de DLBP (Digital Bibliography & Library Project) y de Europeana (figura 2).

Figura 2. CoroArchivo y la tecnología OAI2LOD (Haslhofer y Schnadl, 2008)



3.3. Funcionamiento de la plataforma

El sistema está operativo en la Universidad Central de las Villas (UCLV), donde el convenio de universidades flamencas VLIR tiene instaladas las infraestructuras tecnológicas necesarias para su funcionamiento (tabla II).

A continuación se detalla el proceso de transformación de los datos:

- Paso 1. Extracción de metadatos: haciendo uso de Metharto se localizan los SPARQL Endpoints de los fondos de archivo y se generan las tablas relacionales con los datos para los diversos niveles de descripción ISAD-G y los elementos de SKOS.
- Paso 2. Pre-procesamiento: los datos que se encuentran en DSpace, independientemente de si ya han sido transformados o no, se procesan para llevar a cabo el control de autoridades y que tengan equivalencia con la estructura OAI-PMH. Por medio de reglas preestablecidas se detectan de forma automática los autores duplicados, se asignan títulos uniformes para documentos y publicaciones, y descriptores de materia dentro del entorno de Linked Data.

  Los datos se modelaron utilizando ontologías específicas para la archivística. En el caso de registros bibliográficos de otro tipo de materiales es posible usar Fabio (Peroni y Shotton, 2012) o Bibo (Giasson y D'arcus, 2009), ambas de aplicación dentro del ámbito bibliotecario. En nuestro caso se recurrió a una ontología en formato OWL creada ad hoc por Álvarez (Ledesma, 2013) siguiendo las propuestas de terceros (Sánchez Alonso y otros, 2008). Los elementos de materia se estructuraron con SKOS (Solomou y Papatheodorou, 2010), y las entradas de autoridad de acuerdo con FOAF (Brickley y Miller, 2010).
- Paso 3. Transformación: con el editor de ontologías Protégé se exportaron los datos a DSpace, y con 2RQ se transformaron las bases de datos relacionales en grafos RDF usando un script en lenguaje bash que permite incorporarlos a Virtuoso. En 2RQ existe una herramienta denominada D2R que facilita la transformación de las consultas SPARQL en SQL, y su posterior ejecución sobre las bases de datos relacionales, lo cual es posible cargando y creando un grafo RDF externo.
- Paso 4. Enlace de datos: empleando Silk se establecieron enlaces RDF que apuntaban a fuentes de datos de archivo. De este modo cada conjunto de datos externos quedó accesible para las aplicaciones en forma de triple store. Los datasets VIAF, DLBP y Europeana se descargan en Virtuoso. Disponibles ya en formato SPARQL se configuró Silk para generar enlaces de exclusión e inclusión, realizando el control de autoridades en los datasets con reglas de decisión. Es decir, de forma diferente a otros proyectos comentados en el apartado 2.2, que usan distancias geométricas para establecer clústers de coincidencia o similaridad. Este proceso finaliza con la transformación de los datos a formato LIAM.
- Paso 5. Publicación: los grafos RDF quedan disponibles en la web para los usuarios y para cualquier otro sistema de información. Se decidió usar Pubby (Cyganiak y Bizer, 2008) con SPARQL, ya que rescribe la URI base, a diferencia de SPARQL, que lanza las consultas contra los RDF subyacentes. Como interfaz para la recuperación de información se escogió Drupal, que da soporte a toda esta tecnología.

Tabla II. Conjunto de repositorios de la UCLV

Nombre del Repositorio	OAI-PMH Endpoint
Fondo Coronado Fotos	http://fotos.coroando.uclv.edu.cu/oai
Fondo Coronado Cartogramas	http://cartogramas.uclv.edu.cu/oai
Manuscritos Coronado	http://manuscritos.uclv.edu.cu/oai
Carteles Coronado	http://carteles.uclv.edu.cu/oai
Revista Coronado	http://revistas.uclv.edu.cu/oai
Archivo Histórico Universitario	http://archiuniv.uclv.edu.cu/oai
Archivo de Central de la Universidad	http://centralarchv.uclv.edu.cu/oai

3.4. Módulos de búsqueda semántica y facetada

Drupal, desarrollado a partir de 1999 por Dries Buytaert, y publicado bajo licencia GNU dos años más tarde, genera sitios web combinando varios "bloques" para adaptar sus funcionalidades a las necesidades específicas. Es además un "content management framework" (Byron y otros, 2012). La información se almacena en una base de datos relacional (MySQL, PostgreSQL, SQLite…) por medio del lenguaje de programación PHP. Permite la publicación de datos en formato RDF, además de soportar otros, como n-triples, JSON, XML, RSS y turtle. Gestiona las URIs de los materiales RDF publicados y un SPARQL Endpoint para la consulta. Se pueden personalizar los campos RDF y los namespaces (Alonso-Sierra y otros, 2012).

La búsqueda facetada y las consultas en SPARQL se realizan por medio de los siguientes módulos de Drupal:

- RDF Extensions 7.x-2.0-alpha1
- SPARQL 7.x-2.0-alpha1
- Views 7.x-3.0-beta3
- SPARQL Views 7.x-2.0-alpha2
- Entity API 7.x-1.0-beta8
- CTools 7.x-1.0-alpha4
- SPARQL Views
- RDF UI
- JSONP SPARQL
- OSF (Open Semantic Framework)
- Views UI modules
- OWL2Tips
- DSpace

Instalados los módulos necesarios, OWL2tips (Domínguez-Velasco, 2013a) facilita la visualización en formato OWL y la transformación de las descripciones archivísticas a estructura LIAM, capaz de operar con grandes conjuntos de datos.

3.4.1. Búsqueda federada

Para generar las búsquedas facetadas de ontologías en Drupal se utiliza una versión modificada de la propuesta de Coppens (Coppens y otros, 2013) junto a OWL2tips (Domínguez-Velasco, 2013a). El gran problema de Drupal para la gestión de ontologías es su apego a vocabularios específicos de tratamiento de datos. Lenguajes de ontologías como SKOS, Dublin Core o SIOC son ineficientes cuando se trata de documentación archivística y de búsqueda de información federada. Hay muchos elementos semánticos que provienen del lenguaje OWL que quedan fuera de las consultas clásicas de SPARQL, lo que impide la elaboración de consultas más sofisticadas que exploten las posibilidades de la transitividad y la simetría cuando se trata de localizar información en varios recursos.

El módulo de búsqueda utiliza los elementos OWL y los enlaces de igualdad y de exclusión del lenguaje para establecer nuevas consultas dirigidas a puntos específicos de la estructura OWL, de manera que se pueda acceder a diferentes recursos de información en bases de datos remotas. OWL2tips usa SPARQL Enpoints remotos capaces de enlazar las consultas con cada dependencia del archivo de la Universidad, de la DBLP, VIAF y Europeana. Existen muchos problemas cuando se intenta realizar búsquedas federadas a partir de vocabularios específicos y en repositorios diferentes, pues:

- Cada recurso necesita un SPARQL Endpoint que permita construir relaciones dentro de él.
- Cuando mapeamos un vocabulario no se puede usar una misma consulta para recuperar información en recursos diferentes.
- Es imprescindible trabajar con la especificidad en las consultas y no utilizar elementos genéricos en la descripción, por ejemplo foaf:name en vez de RDFs.
- Las consultas no son efectivas sino rígidas, y obligan a que los usuarios empleen lenguajes de consulta que no conocen y que les hacen complejo localizar la información.
- Los usuarios no pueden recuperar toda la información que requieren, y sus consultas se hacen ineficientes porque no abarcan elementos que la búsqueda clásica de SPARQL es incapaz de gestionar.

Como solución hemos utilizado los enlaces que proceden del lenguaje OWL y constructores que generan un índice de búsqueda que indexa cada elemento de las ontologías. Para esto se almacenan los fondos documentales en los Service Endpoint, y las propiedades de mapeo de los elementos junto a los mapas de cada clase. La información se extrae de las consultas realizadas en lenguaje SPARQL con ayuda de los constructores utilizados en la definición de la ontología, combinados con las propiedades *OWL:same as*, *OWL:equivalentClass*, *OWL:equivalentProperty*, *OWL:differentFrom* y *OWL:AllDifferent* como enlaces. El proceso de indexación está precoordinado para facilitar la recuperación de información, con funcionalidades preestablecidas que enlazan los recursos. Gracias al procesador distribuido de SPARQL es posible

utilizar información pre-indexada para las nuevas consultas que entran en la plataforma y distribuir búsquedas derivadas de consultas generales de acuerdo con las necesidades de los usuarios potenciales del sistema. Esto se consigue recurriendo a la extensión ARQ del módulo JSONP SPARQL de Drupal, con tres procedimientos que permiten consultas federadas: 1) la generación del índice de distribución para procesar los enlaces que describen igualdad y desigualdad de clases; 2) la distribución de las consultas (usando un index building) que facilita la búsqueda federada y el mapeo de las búsquedas a través de índices distribuidos; y 3) la generación de asociaciones de clases e instancias a través de mecanismos de aprendizaje difusos insertados en la herramienta OWL2tips (figura 3).

Figura 3. Segmento de una sentencia SPARQL en CoroArchivo

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX arch: <http://archive.uclv.edu.cu/localontologynamespace#>
CONSTRUCT {?resource owl:sameAs ?remoteresource}
WHERE {
    ?resource a owl:thing.
    ?resource arch:type "person".
    ?resource arch:name ?concept.
    ?concept skos:prefLabel ?name.
   ?concept skos:Concept ?name.
    SERVICE <http://archiveuclv.org/sparql>
    {
      ?remoteresource a foaf:Person.
     ?remoteresource foaf:name ?remotename.
      FILTER (str(?remotename) = ?name)
  }
     }
```

OWL2tips trabaja con las relaciones de diferenciación e igualdad descritas en las ontologías para distribuir las búsquedas y extraer los mapas RDF de cada estructura dentro del módulo JSONP SPARQL de Drupal. Se hacen tres usos concretos de OWL2tips en esta fase:

- Para enriquecer la indexación junto a SPARQL: de esa manera es posible generar consultas que representan el origen de las relaciones de tipo *OWL:same as*, *OWL: equivalentClass*, *OWL:equivalentProperty*, *OWL:differentFrom* y *OWL:AllDifferent*, usadas como medio para hacer enlaces en el dataset. Esto lo logramos mediante constructores para facilitar enlaces internos dentro del dataset. Las consultas son descritas y procesadas con el álgebra SPARQL, y expresadas mediante el algoritmo que maneja OWL2tips que se ha diseñado para trabajar con Drupal específicamente en este proyecto.
- Para complementar los índices de OWL2tips con Silk: creando un archivo con las especificaciones para generar enlaces dentro de Silk y empleando un sistema de indexación capaz de reconocer el origen de los enlaces SPARQL por medio de reglas o constructores de consultas. El sistema de indexación de Silk lleva a cabo un parsing del archivo de configuración de la herramienta Silk y crea los índices mediante los Endpoints de SPARQL, los mapas de clases y las propiedades específicas de los SPARQL Endpoints disponibles directamente en el archivo de configuración para cada una de las propiedades.
- Para aproximar las consultas valiéndose de las posibilidades que ofrecen los enlaces de las propiedades *OWL:same as*, *OWL:equivalentClass*, *OWL:equivalentProperty*, *OWL:differentFrom* y *OWL:AllDifferent*, utilizando algoritmos de aprendizaje basados en lógica difusa, con lo que se logra recuperar con más precisión los llamados conjuntos de datos intermedios, que muchas veces se solapan cuando se recupera la información (figura 4).

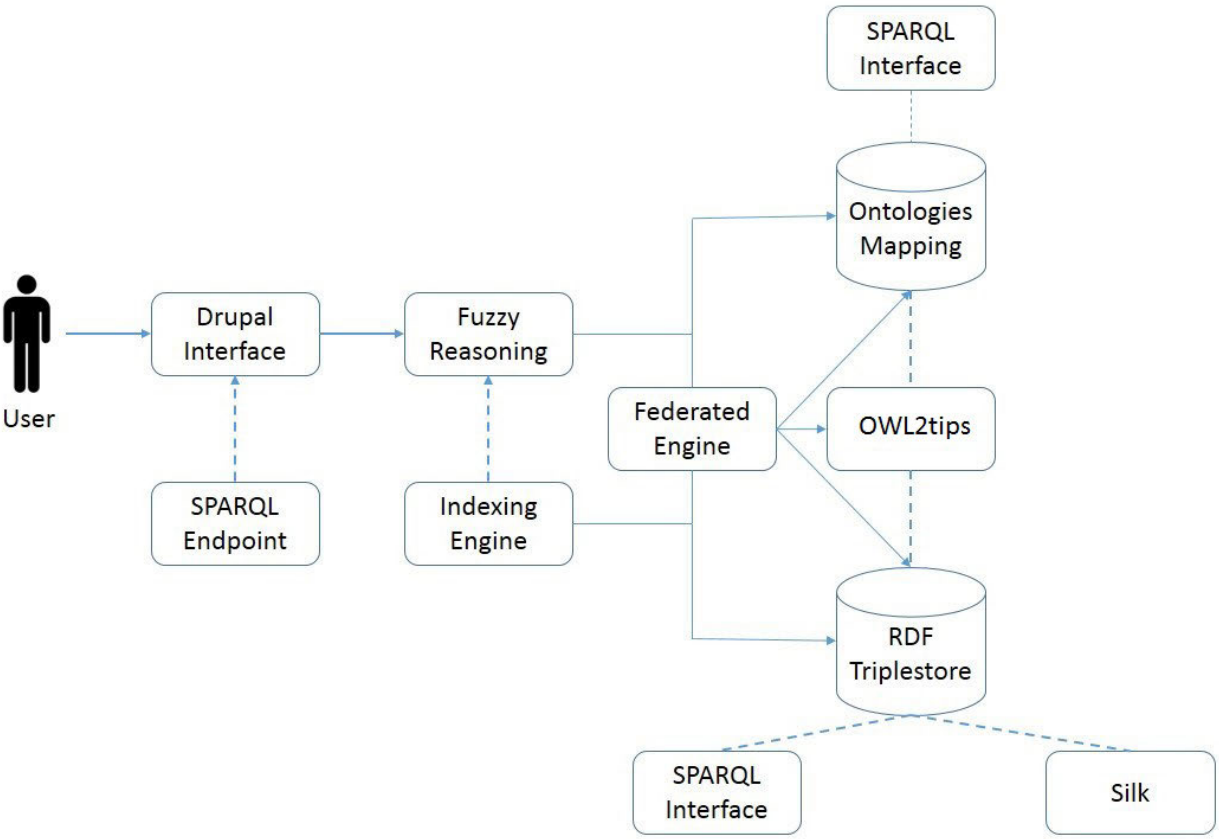
4. PROCESO DE EVALUACIÓN

4.1. El dataset de entrada

Coroimagen UCLV es un conjunto de datos que posee 1.422 instancias, más de 68 datasets y 5 servicios de archivo descargados de VIAF, Europeana, DLBP y de los archivos de la propia institución donde se realiza la prueba. La mayor complejidad para manejar estos recursos es establecer la correspondencia entre los datasets con las fuentes de información, como sucede con DSnotify (Popitsch y Haslhofer, 2011).

Se creó, por ese motivo, un archivo que establece la relación entre los elementos utilizados en la descripción que hace posible medir el nivel de eficiencia de CoroArchivo a partir de los eventos que registra la aplicación y su influencia en f-mesure. Dicho archivo sirvió también para detectar que existía un elevado número de instancias, 98, que estaban mal escritas. Para subsanar este error se crearon indicadores de exclusión que aislaran del experimento aquellas instancias cuyo nivel de ocurrencia afectaba al proceso de evaluación. Así fue posible lograr una distribución homogénea de eventos en un tiempo determinado y minimizar los problemas de entropía.

Figura. 4. Sistema de búsqueda federada de CoroArchivo



4.2. Test

Se decidió realizar la evaluación de acuerdo con los procedimientos de Popitsh y Haslhofer (Popitsch y Haslhofer, 2011), que verifica el funcionamiento de una plataforma con información que implica control de autoridades. En nuestro caso queríamos analizar en qué medida la precisión y la exhaustividad (F-measure) influían en el rendimiento de la herramienta. Para ello se utilizó: un vocabulario capaz de almacenar la información del proceso (Leiva-Mederos y otros, 2013); un módulo de monitorización de eventos; un simulador encargado de representar los datasets en otros entornos de la Universidad; y un analizador que mostraba y examinaba los parámetros seleccionados para emitir los resultados de la evaluación.

Con un algoritmo denominado jump (Domínguez-Velasco y Leiva-Mederos, 2013) se registraron los valores de entropía de las propiedades de los namespaces que se alojan en la plataforma con una cobertura menor del 20%. Estos valores se originaban en caso de que los recursos involucrados en el proceso tuvieran cubiertos sus datasets con al menos un 20% de instancias. Los recursos evaluados fueron Europeana, VIAF, DLBP y el archivo de la UCLV, seleccionando los elementos que caracterizan un documento que tienen buena representación en los datasets:

$$H(p) = -\sum_{i=1}^{n} p_i \ln(P_i)$$

La evaluación de la entropía sirve para seleccionar los namespaces con mejores valores de C, H y Hnorm, de acuerdo con las características de un documento (tabla III). Los namespaces evaluados fueron:

http://xmlns.com/foaf/0.1/

http://xmlns.com/foaf/0.1/

http://www.europeana.eu/schemas/ese/

http://purl.org/dc/elements/1.1/

http://viaf.org/viaf/data/

Tabla III. Cobertura, entropía y entropía normalizada para los datasets de este experimento con un 20% de cobertura

Nombre	Cobertura	H	Hnom
skos:Concept	0,999	0,8345	0,999
skos:prefLabel	0,876	0,8523	0,876
skos:altLabel	0,981	0,8698	0,981
skos:broader	0,976	0,8765	0,976
skos:narrower	0,956	0,8773	0,956
skos:related	0,888	0,3458	0,888
skos:broadMatch	0,947	0,3456	0,947
skos:narrowMatch	0,876	0,8765	0,876
skos:relatedMatch	0,999	0,7772	0,999
skos:exactMatch	0,976	0,7834	0,976
skos:closeMatch	0,978	0,6783	0,978
skos:note	0,999	0,9865	0,999
skos:notation	0,913	0,4531	0,913
skos:inScheme	0,999	0,7803	0,999
rdf:ResourceOrLiteralType	0,991	0,2456	0,991
rdf:LiteralType	0,998	0,7735	0,998
edm:currentLocation	0,965	0,8456	0,965
edm:hasMet	0,953	0,3903	0,953
edm:hasType	0,975	0,7931	0,975
edm:incorporates	0,876	0,563	0,876
edm:isDerivativeOf	0,981	0,455	0,981
edm:isNextInSequence	0,976	0,7653	0,976
edm:isRelatedTo	0,68	0,3457	0,68
edm:isRepresentationOf	0,45	0,2345	0,45
edm:isSimilarTo	0,67	0,3455	0,67
edm:isSuccessorOf	0,98	0,3454	0,98
edm:realices	0,76	0,3467	0,76
edm:EdmType	0,34	0,45678	0,34
edm:ProvidedCHOType	0,67	0,6789	0,67
edm:WebResourceType	0,7	0,5678	0,7
edm:AgentType	0,85	0,345	0,85
edm:PlaceType	0,45	0,567	0,45
edm:TimeSpanType	0,67	0,23	0,67
edm:aggregatedCHO	0,45	0,345	0,45
edm:collectionName	0,78	0,64	0,78
edm:country	0,89	0,71	0,89
edm:hasView	0,678	0,57	0,678
edm:isShownBy	0,89	0,84	0,89
edm:preview	0,78	0,63	0,78
edm:landingPage	0,56	0,4123	0,56

Nombre	Cobertura	H	Hnom
edm:language	0,45	0,33	0,45
edm:rights	0,99	0,93	0,99
owl:same as	0,99	0,93	0,99
dc:contributor	0,99	0,93	0,99
dc:description	0,76	0,7	0,76
dc:title	0,99	0,95	0,99
dc:creator	0,99	0,92	0,99
foaf:document	0,99	0,89	0,99
foaf:mane	0,96	0,91	0,96
owl:differentFrom	0,99	0,9123	0,99
uclvf:title	0,99	0,85	0,99
uclvf:topic	0,96	0,89	0,99
uclvf:type	0,86	0,84	0,86

El resultado de la tabla III muestra que los mejores resultados en el proceso de evaluación de entropía, H y Hn son superiores en los datasets de Europeana y en los de VIAF, debido a la calidad de las descripciones de los datos. Los datasets locales, etiquetados con la sigla uclvf, mantienen buenos índices de evaluación, todos por encima de 0,5. Se decidió usar en el test las características *uclvf:type*, *skos:Concept*, *edm:collectionName*, *foaf:document*, *foaf:mane*, *uclvf:title*, *edm:EdmType*, que tienen mayor representatividad en cada uno de los sistemas de información seleccionados, y que representan lo esencial de un documento.

A continuación se llevó a cabo una simulación, poniendo en funcionamiento el software durante 60 segundos para acceder al 30% de las instancias, como en trabajos que pretendían evaluaciones similares (Sing-Borrajo, 2013), para así monitorizar los cambios en los datasets con una tasa media de 7,03 eventos, resultado del cociente 422/60 seg. Se trataba de analizar la influencia de los eventos de la plataforma en su exhaustividad y precisión. Para ello se repitió la prueba con diferentes intervalos (4 s., 8 s., 15 s., 25 s., 45 s.), que registraron un promedio de 12, 21, 34, 43 y 379 por cada segundo de ejecución del sistema.

Evaluación de la exhaustividad y la precisión

La exhaustividad registra valores muy favorables para el dataset de CoroArchivo, denominado *uclvf:title* y *uclvf:type*. Los porcentajes de estos valores se mantienen por encima del 80%. A medida que aumentan los eventos tienden a mejorar los valores de exhaustividad. Los niveles inferiores de exhaustividad solo aparecen en algunos eventos cuando la plataforma comienza a trabajar. Los elementos de Europeana y VIAF muestran resultados favorables de exhaustividad a medida que se van produciendo eventos. Los tipos de datos que mayor exhaustividad alcanzan a medida que aumentan los eventos son los de DLBP (Figura 5).

Precisión

La precisión también crece a medida que aumentan los eventos. Las etiquetas de Europeana y uclvf registran el mayor nivel de precisión. Pero, si bien es cierto que tiende a acrecentarse a medida que aumentan los eventos, en la mayoría de los namespaces analizados se mantiene por debajo del 40%. Aparecen mayores índices de precisión en los *foaf:name* y *foaf:document* pertenecientes a DLBP, a *uclvf title* de la base CoroArchivo y a *edm:Edmtype*: de Europeana (Figura 6).

F-measure

La figura 7 muestra cómo disminuyen de manera general los valores de F-measure a medida que aumenta el tiempo de trabajo de la plataforma CoroArchivo. Esto no obliga a usar un mayor número de vectores de rasgos para lograr mayores niveles de exhaustividad y precisión, pues ha de tenerse en cuenta que la herramienta procesó datasets de gran tamaño y con altos niveles de información, más de 50 gigas de datos. Por lo tanto los resultados de F-measure parecen aceptables. Es importante destacar que, entre el octavo y el décimo evento, los valores de exhaustividad y precisión (F-measure) alcanzaron resultados muy cercanos a 100%, ya que el servidor tiende a recargarse en los primeros minutos. Con el paso del tiempo mejora su capacidad.

Figura 5. Influencia del número de eventos del sistema CoroArchivo en la exhaustividad

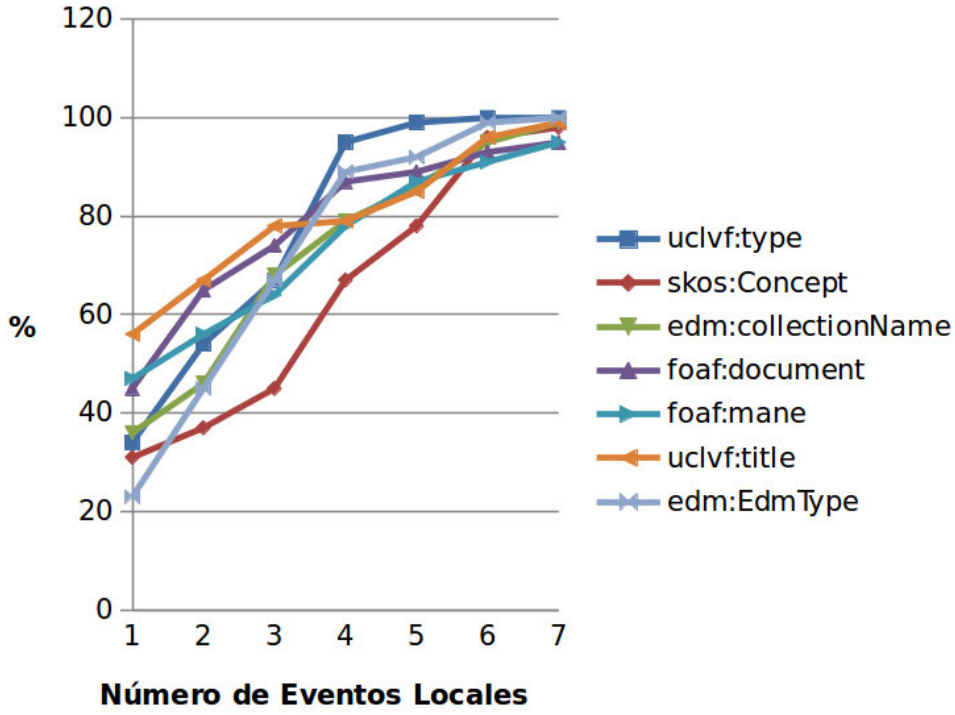


Figura 6. Influencia de los eventos locales de la plataforma en la precisión del sistema

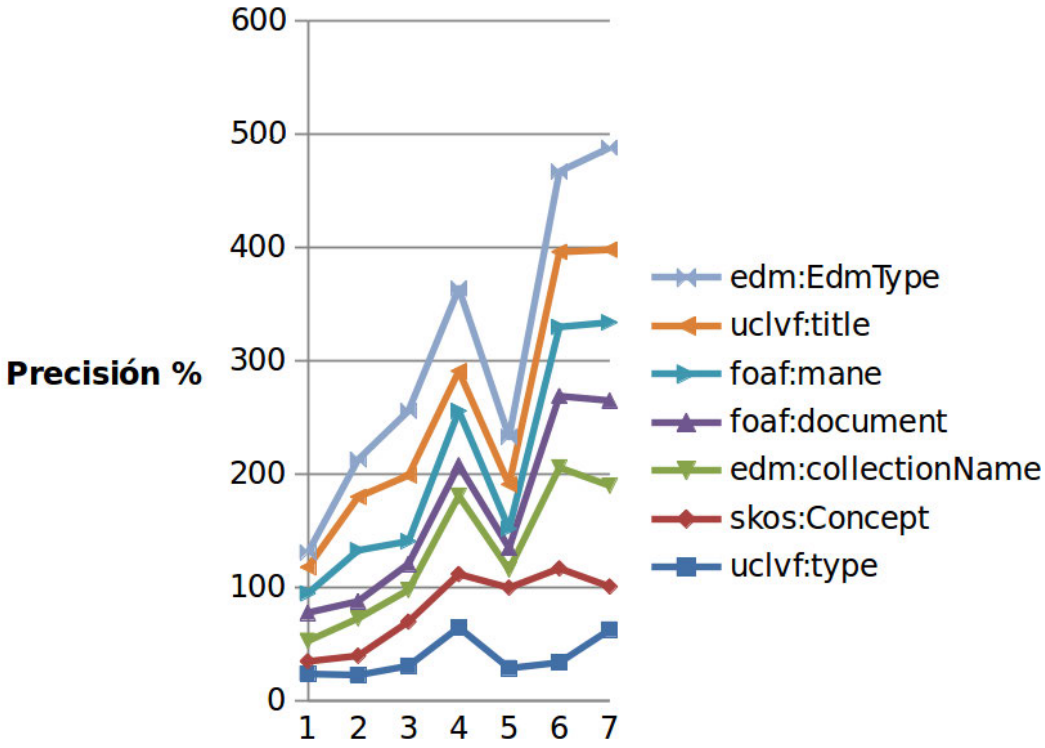
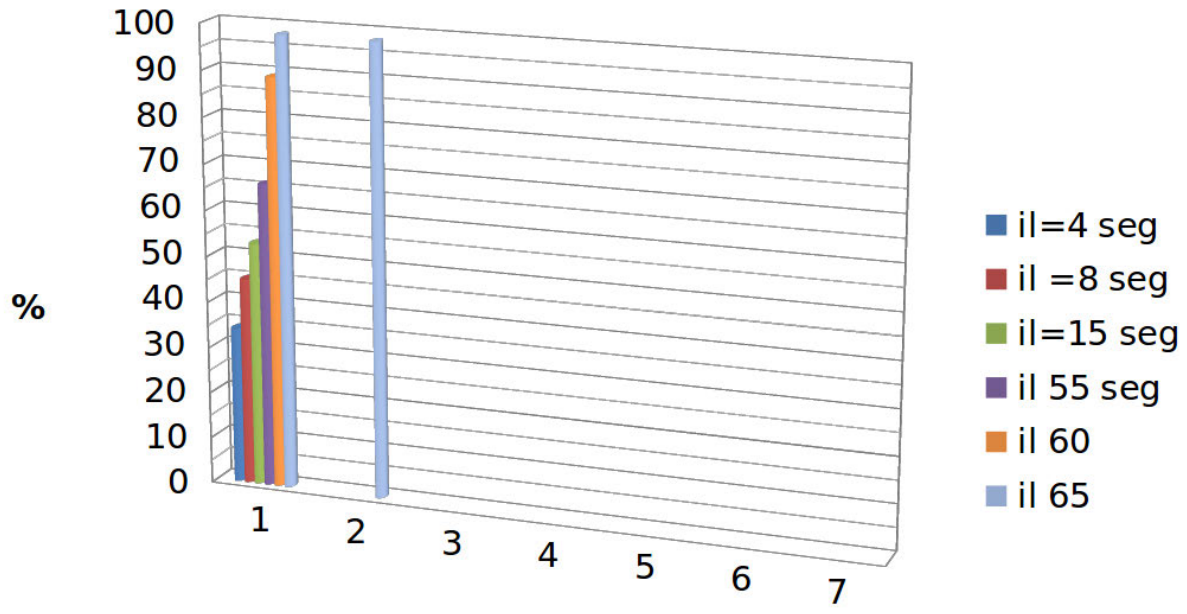


Figura 7. Influencia de los eventos locales de la plataforma en el parámetro F-measure con variables de tiempo



Rendimiento en la búsqueda federada

Para evaluar la búsqueda federada usamos el SPARQL Endpoint interno de los fondos de archivos de la UCLV, dataset con 30.000 registros de archivo publicados simultáneamente en un SPARQL Endpoint local y en otro remoto. En el local solo se almacenan las URI de los recursos, mientras que en el remoto se incluyen recursos de información con propiedades ontológicas específicas para igualar y excluir elementos. La búsqueda en el SPARQL Endpoint local y en el remoto facilita la consulta de información federada y la evaluación de OWL2tips. Se analizó la consulta distribuida de excelencia, realizada a través de la plataforma CoroArchivo y su SPARQL Endpoint, ponderando la eficiencia de la herramienta en función del tiempo y el número de consultas generadas con el algoritmo. Otra parte del proceso de validación de la herramienta realizó el mismo procedimiento en otros datasets (Europeana, DLBP y VIAF) descargados en el servidor de la UCLV, con 68.000 registros, a los que se le adjudicaron constructores y modelos de etiquetado similares a los de CoroArchivo en OWL (tabla IV).

La evaluación de la efectividad del sistema de búsqueda federada a través de las capacidades de su algoritmo demuestra que la herramienta logra valores muy positivos cuando se realizan las búsquedas sobre los recursos internos de la UCLV. Los primeros resultados de las consultas con el algoritmo global aportan más del 30% de los datos alojados en los recursos semánticos, tanto durante los primeros momentos de la recuperación de la información como en el final del proceso, momento en el que aumenta el nivel de exhaustividad hasta el 97% de los recursos con instancias. La valoración de la búsqueda distribuida perfecta también mostró resultados favorables durante este proceso. Cuando se realizó el mismo experimento con los datos de otras fuentes como Europeana, DLBP y VIAF, descargados en los servidores de la UCLV, los resultados fueron similares, y presentan mayores niveles de recuperación de información. Sin embargo aún existen deficiencias en la descripción, que ha dejado muchos recursos de archivo de valor bibliográfico mal descritos y, en consecuencia, no recuperados (tabla V).

5. CONCLUSIONES

Tras analizar las principales tecnologías aportadas por la Web Semántica, algunos de los proyectos que las están implementando para la gestión de fondos de archivos, así como la plataforma que se ha desarrollado específicamente para el presente trabajo, queda manifiesto, a nuestro juicio, que resulta beneficioso utilizar dichas tecnologías para gestionar y recuperar informaciones de archivo con datos enlazados.

La puesta en funcionamiento de este tipo de sistemas puede hacer uso de diversas herramientas ya existentes, como Silk, Virtuoso, Pubby, Authoris y Metharto, que facilitan la generación de modelos de datos enlazados de alto nivel semántico.

Tabla IV. Efectividad de la búsqueda federada en el dataset local

Algoritmo de búsqueda perfecta	Primeros resultados de tiempo	Últimos resultados de tiempo
Búsqueda 1	1023 ms	13456 ms
Búsqueda 8	4213 ms	14233 ms
Búsqueda 10	6783 ms	16234 ms
Uso global del algoritmo	**Primeros resultados de Tiempo**	**Últimos resultados de tiempo**
Búsqueda 1	12034 ms (+41%)	22116 ms (+73%)
Búsqueda 8	19076 ms (+63%)	29123 ms (+97%)
Búsqueda 10	10023 ms (+33,41%)	29113 ms(+97%)

Tabla V. Efectividad de la búsqueda federada en los datasets de Europeana, DLBP y Library of Congress

Algoritmo de búsqueda perfecta distribuida	Primeros resultados de tiempo	Últimos resultados de tiempo
Búsqueda 1	150023 ms	205677 ms
Búsqueda 8	345783 ms	345680 ms
Búsqueda 10	387643 ms	416234 ms
Uso global del algoritmo	**Primeros resultados de tiempo**	**Últimos resultados de tiempo**
Búsqueda 1	386650 ms (56%)	560770 ms (82%)
Búsqueda 8	493050 ms (72%)	602304 ms (88%)
Búsqueda 10	582340 ms (85%)	675143 ms (99%)

El modelo propuesto en este artículo se basa en las prestaciones ofrecidas por las propiedades de los objetos para formular búsquedas federadas, haciendo posibles nuevas opciones para la consulta de conjuntos de datos.

Consideramos, por otra parte, que la experiencia de evaluación incluida en estas páginas demuestra suficientemente la eficiencia del modelo, pues la plataforma logra altos niveles de F-measure, en la mayoría de los casos por encima del 60%.

6. REFERENCIAS

Addis, M.; Allasia, W.; Bailer, W. (2010). 100 Million Hours of Audiovisual Content: Digital Preservation and Access in the PrestoPRIME Project. *First International Digital Preservation Interoperability Framework (DPIF) Symposium*. Dresden, Germany: ACM.

Alonso-Sierra, L. E., Ortiz-Muñoz, E.; Hidalgo-Delgado, Y. (2012) Los sistemas de gestión de contenidos en el ámbito de la web semántica: una breve revisión. *Serie científica de la Universidad de las Ciencias*, 5, 1-9.

Baker, T. (2012). Libraries, languages of description, and linked data: a Dublin Core perspective. Library Hi Tech, 30, pp. 116 - 133. http://dx.doi.org/10.1108/07378831211213256

Berners-Lee, T.; Hendler, J. y Ora Lassila (2001). The Semantic Web. *Scientific American*, vol. 284 (5), 28-37. http://dx.doi.org/10.1038/scientificamerican0501-34

Bizer, C.; Seaborne, A. (2004). D2RQ - Treating Non-RDF Databases as Virtual RDF Graphs. *3rd International Semantic Web Conference*. Hiroshima, Japan: Springer.

Brickley, D.; Miller, L. (2010). *FOAF Vocabulary Specification*.

Byron, A., Berry, A.; De-Bondt, B. (2012). *Using Drupal: choosing and configuring modules to build dynamic websites*. Sebastopol (California), O'Reilly Media.

Coppens, S.; Mannes, E.; Walle, R. V. D. (2009). Disseminating heritage records as linked open data. *International Journal of Virtual Reality*, 8, 39-44.

Coppens, S.; Verborgh, R.; Mannens, E.; Walle, R. V. D. (2013). Querying the Linked Data Graph using owl:sameAs Provenance. *Proceedings of the 16th International Conference on Model Driven Engineering Languages and Systems*.

Cyganiak, R.; Bizer, C. (2008). Pubby - A Linked Data Frontend for SPARQL Endpoints. http://wifo5-03.informatik.uni-mannheim.de/pubby

Doerr, M.; Gradmann, S.; Hennicke, S.; Isaac, A.; Meghini, C.; Van De Sompel, H. (2011). The Europeana data model (EDM). *IFLA 2011: World library and information congress: 76th IFLA general conference and assembly*. Gothenburg, Suecia: IFLA.

Domínguez-Velasco, S. (2013a). *OWL2tips: herramienta para generar y visualizar ontologías desde Drupal*. Beta 2 ed. Santa Clara, Cuba: Universidad Central de las Villas.

Domínguez-Velasco, S. (2013b). *Sampras*. Alfa 2 ed. Santa Clara, Cuba: Universidad Central de las Villas.

Domínguez-Velasco, S.; Leiva-Mederos, A. (2013). *Jump*. Beta ed. Santa Clara, Cuba: Universidad Central de las Villas.

Ferro, N.; Silvello, G. (2013). NESTOR: A formal model for digital archives. *Information Processing and Management*, vol. 49 (6), 1206-1240. http://dx.doi.org/10.1016/j.ipm.2013.05.001

Giasson, F.; D'arcus, B. (2009). *Bibliographic Ontology Specification*. Structured Dynamics LLC. http://bibliontology.com/specification

Görlitz, O.; Staab, S. (2011). SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions. *The 10th International Semantic Web Conference. http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/Workshops/COLD/GoerlitzAndStaab_COLD2011.pdf*

Gracy, Karen F. *(2014). Archival description and linked data: a preliminary study of opportunities and implementation challenges. Arch Sci.* Disponible en *http://doiI10.1007/s10502-014-9216-2*

Grimoüard, Claire Sibille-de (2014). The Thesaurus for French Local Archives and the Semantic Web. Procedia - *Social and Behavioral Sciences, (147), 206–212.* http://dx.doi.org/10.1016/j.sbspro.2014.07.153

Hartig, O.; Bizer, C.; Freytag, J. D. (2009). Executing SPARQL Queries over the Web of Linked Data. *ISWC '09 Proceedings of the 8th International Semantic Web Conference*. Springer-Verlag. http://dx.doi.org/10.1007/978-3-642-04930-9_19

Haslhofer, B.; Roochi, E. M.; Schandl, B.; Zander, S. (2011). *Europeana RDF Store Report*. Viena, Austria: Universidad de Viena.

Haslhofer, B.; Schnadl, B. (2008). The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data. In: *International Workshop on Linked Data on the Web* (LDOW2008). Beijing, China.

Haslhofer, B.; Schnadl, B. (2010). Interweaving OAI-PMH data sources with the linked data cloud. *International Journal of Metadata, Semantics and Ontologies*, vol. 5 (1), 17-31. http://dx.doi.org/10.1504/IJMSO.2010.032648

Hernández, A. (2007). *Organización y representación del conocimiento: paradigmas, hipertextos y fundamentación metamodélica*. Universidad de La Habana. Cuba.

Hidalgo-Delgado, Y.; Rodríguez-Puente, R.; Ortiz-Muñoz, E.; Alonso-Sierra, L. E. (2013). Herramienta para la recolección de metadatos bibliográficos mediante el protocolo OAI-PMH. II *Conferencia Internacional de Ciencias Computacionales e Informáticas*. La Habana, Cuba.

Jentzsch, A.; Isele, R.; Bizer, C. (2010). Silk - Generating RDF Links while publishing or consuming Linked Data. *International Semantic Web Conference Posters&Demos*, Shanghai, China.

Koutsomitropoulos, D. A.; Solomou, G. D.; Theodore S. Papatheodorou (2008). Semantic Interoperability of Dublin Core Metadata in Digital Repositories. *Innovations in Information Technology, 2008. IIT 2008. International Conference on IEEE*. http://dx.doi.org/10.1109/INNOVATIONS.2008.4781709

Ledesma, G. A. (2013). *Coroimagen. Ontología para el tratamiento de archivo*. Santa Clara, Cuba: Universidad Central de las Villas.

Leiva-Mederos, A.; Senso, J. A.; Domínguez-Velasco, S.; Hípola, P. (2013). Authoris: a tool for authority control in the Semantic Web. *Library Hi Tech*, vol. 31 (3), 536-553. http://dx.doi.org/10.1108/LHT-12-20112-0135

Lynch, Tom J. (2014). Social Networks and Archival Context Project: A Case Study of Emerging Cyberinfrastructure. *Digital Humanities Quarterly*, 8(3). Disponible en: http://www.digitalhumanities.org/dhq/vol/8/3/000184/000184.html [Consulta: 1 de septiembre de 2015].

Mena, M. (2006). *Retos de la actividad archivística: reporte de conferencias*. La Habana, Cuba: Universidad de la Habana.

Merlino-Santesteban, C. (2012). Repositorios institucionales y buscadores web: una interrelación no tan exitosa. *10ª Jornada sobre la biblioteca digital universitaria*. Buenos Aires. Argentina.

Moyano Collado, Julián. (2013). La Descripción Archivística. de los Instrumentos de Descripción Hacia la Web Semántica. *Anales de Documentación,* 16(2), 3–13.

Nam-Park, Ok. (2015). Development of Linked Data for Archives in Korea. *DLib Magazine*, 21(3-4), 1-13.

Nolle, A.; Nemirovski, G. (2013). ELITE: An Entailment-based Federated Query Engine for Complete and Transparent Semantic Data Integration. In Eiter, T.; Glimm, B.; Kazakov, Y.; Krötzsch, M. (eds.) (2013). *26th International Workshop on Description Logics*. Ul, Alemania.

Peroni, S.; Shotton, D. (2012). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 17, 33-43. http://dx.doi.org/10.1016/j.websem.2012.08.001

Popitsch, N.; Haslhofer, B. (2011). DSNotify - A solution for event detection and link maintenance in dynamic datasets. *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 9 (3), 266-283. http://dx.doi.org/10.1016/j.websem.2011.05.002

Quilitz, B.; Leser, U. (2008). Querying Distributed RDF Data Sources with SPARQ. *ESWC'08 Proceedings of*

the 5th European semantic web conference on The semantic web: research and applications. Springer-Verlag.

Rademaker, A.; Borges-Oliveira, D. A.; de Paiva, V.; Higuchi, S.; Medeiros e Sá, A.; Alvim, M. (2015). A linked open data architecture for the historical archives of the Getulio Vargas Foundation. *Int J Digit Libr*, (15), 153–167. http://dx.doi.org/10.1007/s00799-015-0147-1

Sánchez Alonso, S.; Sicilia Urbán, M. Á.; Rato Leguina, G. D. (2008). Sobre la interoperabilidad semántica en las descripciones archivísticas digitales. *Revista Española de Documentación Científica*, vol. 31 (1), 11-34.

Schöpfel, J.; Bescond, I.; Prost, H. (2012). Open is not enough: a case study on grey literature in an OAI environment. *The Grey Journal*, vol. 8 (2), pp. 112-124.

Schwarte, A.; Haase, P.; Hose, K.; Schenkel, R.; Schmidt, M. (2011). FedX: A Federation Layer for Distributed Query Processing on Linked Open Data. *The Semanic Web: Research and Applications,* 481-486. http://dx.doi.org/10.1007/978-3-642-21064-8_39

Sing-Borrajo, P. (2013). *Reporte de cargas en los datasets de la biblioteca universitaria de la Universidad Central de las Villas*. Santa Clara, Cuba: Universidad Central "Marta Abreu" de las Villas, Facultad de Ingeniería Eléctrica, Departamento de Telecomunicaciones.

Solomou, G.; Papatheodorou, T. (2010). The Use of SKOS Vocabularies in Digital Repositories: the DSpace case. *Fourth International Conference on Semantic Computing*. IEEE. http://dx.doi.org/10.1109/icsc.2010.83

Tummarello, G.; Delbru, R.; Oren, E. (2007). Sindice.com: Weaving the open linked data. *6th international The semantic web and 2nd Asian conference on Asian semantic web conference.*

Vasallo, S. (2010). Descrizioni Archivistiche e web semántico: un connubio possibile. *Italian Journal of Library and Information Science*, 1(1), 169 – 163.