DE MARC21 A RDF

Mostaccio, María Rosa

Universidad de Buenos Aires. Facultad de Filosofía y Letras. INIBI.

Correo electrónico: mmostaccio@gmail.com

Blanco, Nancy

Universidad de Buenos Aires. Facultad de Filosofía y Letras. INIBI.

Correo electrónico: nancybl@filo.uba.ar

Gattafoni, Silvia

Universidad de Buenos Aires. Facultad de Filosofía y Letras. INIBI.

Correo electrónico: sgattafoni@filo.uba.ar

Uviña, Ramiro

Universidad de Buenos Aires. Facultad de Filosofía y Letras. INIBI.

Correo electrónico: ramirouvia@gmail.com

Duranti, Bárbara

Universidad de Buenos Aires. Facultad de Filosofía y Letras. INIBI.

Correo electrónico: durantibarbara@gmail.com

Yedid, Nadina

Universidad de Buenos Aires. Facultad de Filosofía y Letras. INIBI.

Correo electrónico: nadineyedid@hotmail.com

Resumen: El presente trabajo introduce los aspectos principales de la investigación de “De MARC21 a RDF”. Propone indagar sobre métodos de transliteración que permitan convertir los datos de los catálogos en línea de acceso público (OPAC) de las bibliotecas en formato MARC21 al modelo estándar de intercambio de datos RDF, dentro de la web semántica. Para ello, se repasa brevemente las tecnologías de la web semántica y las interrelaciones existentes entre cada una de ellas.Se da cuenta de la metodología de trabajo que el grupo de investigación está llevando adelante, y se enumeran las principales características de los softwares a considerar.

Introducción

El presente trabajo se propone dar cuenta del proyecto de investigación "De Marc21 a RDF", enmarcado en la línea de investigación tecnológica impulsada por el Instituto de Investigaciones Bibliotecológicas (INIBI), presentado en la convocatoria 2015 de Proyectos de Reconocimiento Institucional de Investigadores Graduados (PRIG) de la Facultad de Filosofía y Letras de la Universidad Nacional de Buenos Aires (UBA).

A lo largo de los años las unidades de información han aplicado estándares que permiten describir y organizar la información de forma tal que los registros bibliográficos obtenidos puedan ser compartidos por diferentes instituciones. El formato MARC21 se ha configurado como uno de los estándares de uso más extendido en las bibliotecas del mundo entero para la descripción de recursos bibliográficos y registros de autoridad. Sin embargo, la estructura del formato MARC21 no es compartida más allá de las bibliotecas, convirtiendo las bases de datos generadas en este formato, en silos de información que no se relacionan de forma alguna con el resto de la información disponible en la web.

La World Wide Web (WWW) se desarrolla en 1989/1990, a partir de la iniciativa llevada adelante por Tim Berners-Lee, ingeniero del Centro Europeo de Física Nuclear (CERN), para crear una herramienta para la búsqueda y transmisión de información entre científicos. Está basada, funcional y estructuralmente, en el hipertexto (hoy hipermedios) y en las redes de computadoras (Lamarca Lapuente, 2006; Coyle, 2012). Actualmente, constituye uno de los servicios de Internet más exitosos, permitiendo la distribución de recursos digitales, conectados mediante hiperenlaces. Su impacto ha provocado un cambio profundo en las maneras de acceder, crear y controlar la información, ya que "la información electrónica presenta una arquitectura hipertextual, no lineal y distribuida que trae consigo un nuevo cambio en el prototipo de la organización del conocimiento" (Méndez Rodríguez, 2002; Daudinot Founier, 2006).

En la Web, la identificación, localización y recuperación de recursos digitales se hace posible a través del uso de metadatos. Estos se incorporan a los recursos mediante lenguajes de marcado, que permiten la codificación de los documentos mediante etiquetas o marcas que contienen información adicional acerca de la estructura, presentación y/o contenido.

Actualmente confluyen en la Web diversas tecnologías que permiten pensar la WWW como una web semántica o web de datos. En base a herramientas como RDF (Resource Description Framework / Marco de Descripción de Recursos), SPARQL (SPARQL Simple Protocol and RDF Query Language / lenguaje estandarizado de consulta para RDF) y OWL (Web Ontology Language / Lenguaje de Ontologías Web), se contribuye a convertir la Web en una

infraestructura global en la que es posible compartir y reutilizar datos y documentos entre diferentes tipos de usuarios.

La presente investigación se encuentra orientada a descubrir y describir métodos a partir de los cuales se posibilite una conversión de la sintaxis del registro MARC21 a una descripción semántica de lo descrito en los registros bibliográficos.

Con este trabajo se espera poder generar alternativas e identificar posibles cursos de acción tendientes a convertir los OPAC de las bibliotecas que se encuentran en silos aislados en la web a recursos disponibles, interrelacionados y accesibles en la web semántica, cuyo desarrollo se encuentra en la agenda internacional del acceso a la información.

Objetivos y alcances

El objetivo del proyecto es conocer los procesos y herramientas para lograr la conversión de registros bibliográficos MARC21 a RDF. Además se pretende proveer un circuito de trabajo que colabore a la toma de decisiones a este respecto.

Tendrá un enfoque cualitativo de carácter exploratorio. Se establecerá una muestra de registros bibliográficos en MARC21. Se seleccionarán diferentes herramientas *Free and Open Source Software* (en adelante FOSS) para convertirlos a RDF. Una vez instalados y probados los sistemas se procederá a plantear un modelo aceptable de circuito de trabajo que permita realizar esta migración en forma ordenada y consistente. Se construirá una herramienta de análisis (*lista de verificación*) para evaluar y comparar diferentes *softwares*.

Metodología

El proyecto de investigación se plantea las siguientes etapas:

1. Mapeo de Marc21 a esquema de metadatos MODS
2. Análisis del contexto de la Web Semántica
3. Identificación y selección de herramientas FOSS para convertirlos de MARC21 a RDF
4. Selección de una muestra de registros bibliográficos en formato MARC
5. Estudio y manejo de los softwares seleccionados
6. Prueba con los conjuntos de datos MARC
7. Elaboración de una matriz de comparación
8. Diseño de una lista de verificación como herramienta de comparación genérica que sirva para estos y otros sistemas

Etapa 1 :Mapeo de Marc21 a esquemas de metadatos MODS

Para poder realizar una conversión desde MARC21 se hace necesario establecer el mapeo de campos desde MARC21 y su posible transliteración a un esquema de metadatos.Según se observa en el siguiente ejemplo, se realizó un mapeo de MARC21 a MODS[1]:

Registro MARC:

```
LDR    00000aam a22 a 4500
001    001281441
005    20150508155824.0
007    ta
008    130531s2012 ec a gr 001 p spa
020    |a 978-9978-48-276-6
040    |a AR-BaBN |b spa |c AR-BaBN |e aacr
0801   |a 821.134.2(82)-1|2 2000
1001   |a Hernández, José, |d 1834-1886
24513  |a El gaucho Martín Fierro / |c José Hernández ; ilustrado por Tito Martínez
; [estudio introductorio por Verónica Jarrín].
250    |a 1a ed.
260    |a Quito, Ecuador : |b Velásquez & Velásquez, |c 2012.
300    |a 166 p. : |b il. ; |c 21 cm.
4900   |a Juvenalia. Literatura latinoamericana |v 48
500    |a Incluye estudio introductorio: p. 7-31.
500    |a Incluye "Actividades que se pueden realizar en el aula": p. 32-33.
60014  |a  Hernández,  José,  |d  1834-1886.  |t  Martín  Fierro  |x  Crítica  e
interpretación
650 4  |a Gauchos |v Poesía
650 4  |a Poesía argentina |y Siglo XIX
655 4  |a Poesía gauchesca |z Argentina |y Siglo XIX
7001   |a Martínez, Tito |e il.
7001   |a Jarrín, Verónica |e autora de la introducción
```

Metadatos Mods:

```
<?xml version="1.0" encoding="UTF-8"?>
<modsCollection xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.loc.gov/mods/v3"
xsi:schemaLocation="http://www.loc.gov/mods/v3
http://www.loc.gov/standards/mods/v3/mods-3-4.xsd">
       <mods version="3.4">
              <titleInfo>
                     <title>El gaucho Martín Fierro</title>
              </titleInfo>
              <name type="personal">
                     <namePart>Hernández, José</namePart>
                     <role>
                            <roleTerm>autor</roleTerm>
                     </role>
```

[1] http://www.loc.gov/standards/mods/mods-mapping.html

```
</name>
<name type="personal">
	<namePart>Martínez, Tito</namePart>
	<role>
		<roleTerm>ilustrador</roleTerm>
	</role>
	</name>
<name type="personal">
	<namePart>Jarrín, Verónica</namePart>
	<role>
		<roleTerm>autora de la introducción</roleTerm>
	</role>
	</name>
<typeOfResource>texto</typeOfResource>
<genre authority=”local”>libro</genre>
<originInfo>
	<place>
		<placeTerm>Quito, Ecuador</placeTerm>
	</place>
	<publisher>Velásquez & Velásquez</publisher>
	<dateIssued>2012</dateIssued>
	<dateCaptured>20150930</dateCaptured>
	<edition>1a. ed.</edition>
</originInfo>
<relatedItem type=”series”>Juvenalia. Literatura
latinoamericana</relatedItem>
<language>
	<languageTerm type=”código”>spa</languageTerm>
</language>
<physicalDescription>
	<extent>159 p.</extent>
</physicalDescription>
<note>Incluye estudio introductorio: p. 7-31.</note>
<note>Incluye "Actividades que se pueden realizar en el aula": p.
32-33.</note>
<subject authority=”unescot”>
	<topic>Gauchos</topic>
</subject>
<subject authority=”unescot”>
	<topic>Poesía</topic>
</subject>
<subject authority=”unescot”>
	<topic>Poesía argentina</topic>
</subject>
<subject authority=”unescot”>
	<topic>Poesía gauchesca</topic>
</subject>
<subject authority=”unescot”>
	<geographic>Argentina</geographic>
</subject>
<subject authority=”unbist”>
	<temporal>Siglo XIX</temporal>
</subject>
	<classification authority=”udc”>821.134.2(82)-1<classification>
```

```
            <identifier type="isbn">978-9978-48-276-6</identifier>
            <location>
                  <physicalLocation>OED</physicalLocation>
            </location>
            <recordInfo>
                  <recordContentSource>AR-BaBN</recordContentSource>
                  <recordCreationDate>20150508</recordCreationDate>
                  <recordIdentifier>001281441</recordIdentifier>
                  <recordOrigin>Reformateado de un registro en
            MARC</recordOrigin>
                  <languageOfCataloging>
                        <languageTerm type="código">spa</languageTerm>
                  </languageOfCataloging>
                  <descriptionStandard>aacr</descriptionStandard>
            </recordInfo>
      </mods>
</modsCollection>
```

Etapa 2 : Análisis del contexto de la Web Semántica

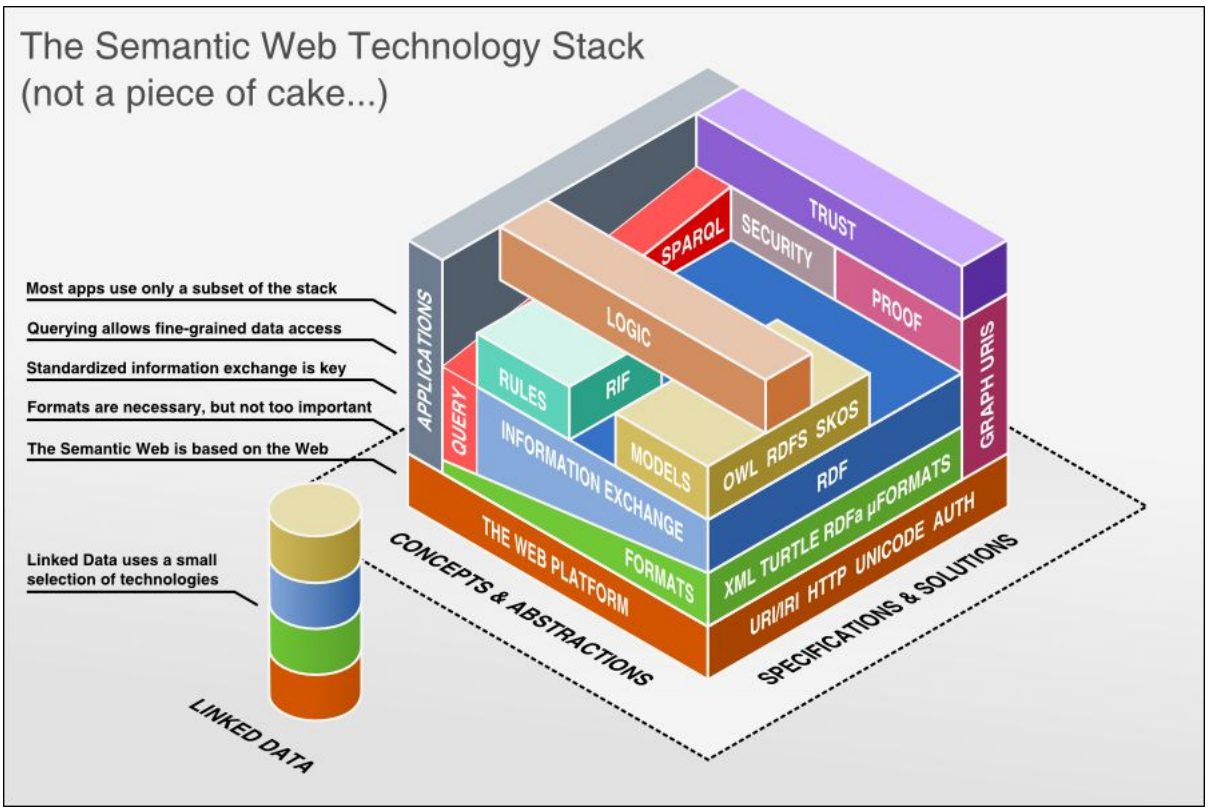
El funcionamiento de la Web semántica integra en la Web actual la posibilidad de agregar varios datos relacionados entre sí, ya sea semánticamente o por atributos que los determinan. Según una definición del W3C "*la web semántica es una web extendida, dotada de mayor significado en la que cualquier usuario en internet puede encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida. Al dotar a la web de más significado, y, por lo tanto, de más semántica, se pueden obtener soluciones a problemas habituales en la búsqueda de la información, mediante la utilización de una infraestructura común, que permite compartir, procesar y transferir información de forma sencilla*". (W3C[2])

"…La WWW actual es básicamente una web de documentos enlazados, mediante hipervínculos que relacionan los recursos unos con otros. Estos enlaces en sí mismos no tienen ningún significado más allá que "enlace", pero no se especifica de forma comprensible ni por las máquinas ni por las personas, qué tipo de enlace se establece entre ambos recursos. En esta Web, los motores de búsqueda localizan los documentos mediante la comparación entre los términos de búsqueda que introduce un usuario con las palabras contenidas en los documentos. Esta forma de búsqueda es bastante distinta a la que se realiza en una base de datos con metadatos, es decir, descripciones estructuradas de los recursos, afirmaciones que tienen algún nivel de significado y que permiten por tanto realizar consultas a la base de datos. El movimiento Linked Open Data pretende emplear metadatos con significado semántico y hacer que sustenten sistemas de búsqueda en un entorno web abierto, comenzando por enlazar datos en lugar de enlazar simplemente documentos. La mayor parte

[2] http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica

de los metadatos actualmente residen en bases de datos, como catálogos de biblioteca, sistemas de biblioteca digital, entre otros. Por lo tanto, los metadatos se localizan en silos de datos aislados que no pueden ser enlazados ni consultados de forma simultánea. Los motores de búsqueda no pueden acceder a estos recursos para indizarlos y sustentar la búsqueda en base a metadatos. Es lo que denominamos comúnmente como Web invisible, o Web profunda. Así, Linked Data consiste en emplear la web para conectar datos relacionados que no estaban enlazados previamente, o usar la Web para minimizar las barreras para enlazar datos que actualmente están enlazados con otros métodos.¨ (Méndez Rodríguez; Bueno de la Fuente, 2014).

Estas tecnologías se representan generalmente como una serie de capas, en la que las especificaciones y estándares se organizan en torno a los conceptos y abstracciones relativos a la plataforma o arquitectura web, los formatos, el intercambio de información, los modelos de datos, la consulta, y otros aspectos clave para la localización, recuperación y acceso a los datos, mediante la definición de reglas de inferencia, la lógica, la seguridad, la protección y la confianza.



Fuente: http://bnode.org/media/2009/07/08/semantic_web_technology_stack.png

Entre las especificaciones fundamentales se encuentra:

Arquitectura Web

En la base de la plataforma web se encuentran los IRIs (*International Resource Identifiers*). Los IRIs son generalizaciones de los URIs (*Uniform Resource Identifiers*) que permiten un espectro más grande de caracteres Unicode, como por ejemplo caracteres no-ASCII. Se puede decir que todos los URI y los URL (*Uniform Resource Locator*) son IRI, pero no todos los IRI son URI o URL. Más allá de la sintaxis permitida en cada uno de estos identificadores, la principal aplicación práctica es que un IRI permite representar recursos que no estén disponibles en la web. Es decir, un IRI no siempre implica una dirección de la web en la cual se puede ir a localizar el recurso, sino que sólo funciona como un identificador unívoco (W3C, 2014).

Formatos

El modelo RDF utiliza formatos que permiten exponer directamente los datos en la web (Styles, Ayers, Shabir, 2008), como documentos o incrustar datos estructurados mediante la especificación de atributos en cualquier lenguaje de marcado.Estos se denominan formatos de serialización (*marshalling*) y, desde el punto de vista informático, permiten "salvar un objeto digital temporalmente en un medio de almacenamiento en una ubicación remota e independiente con objeto de que pueda ser transmitido o almacenado posteriormente en otro lugar" (Voutssas-Márquez, Barnard Amozorrutia, 2014), en otras palabras, la serialización es "un proceso de codificación de un objeto en un medio de almacenamiento (como puede ser un archivo, o un buffer de memoria) con el fin de transmitirlo a través de una conexión en red como una serie de bytes o en un formato humanamente más legible como XML o JSON, entre otros. La serie de bytes o el formato pueden ser usados para crear un nuevo objeto que es idéntico en todo al original, incluido su estado interno (por tanto, el nuevo objeto es un clon del original). La serialización es un mecanismo ampliamente usado para transportar objetos a través de una red, para hacer persistente un objeto en un archivo o base de datos, o para distribuir objetos idénticos a varias aplicaciones o localizaciones" (Colaboradores de Wikipedia, 2015).

De este tipo de formatos, si bien XML es el de uso más extendido, dado que el formato de intercambio de datos puede impactar en el rendimiento y en la velocidad respecto de la transmisión de datos (Nurseitov, Paulson, Reynolds, Izurieta, 2009), actualmente JSON presenta importantes beneficios: "archivos más simples y pequeños; información más legible y fácil de mantener; disponibilidad de utilerías para procesar JSON, menos código y menos errores de programación" (Pacheco, Ramírez, Guzmán, Cruz-Flores, 2013). XML tiende a ser mucho más difícil de manejar para los programadores que JSON, porque no fue diseñado

principalmente para ellos (Quin, 2013). Una diferencia importante en este sentido, es que JSON almacena los datos en vectores y registros mientras que XML los almacena árboles.

En 2014, el W3C, en el documento RDF 1.1 Primer "introduce nuevos formatos de serialización, de modo que en la definición de los conceptos y sintaxis RDF/XML, ya no es el único formato de serialización recomendado" (Garzón Farinós, 2014).

Existen diversas formas de codificación sintáctica para la notación de estas tripletas, ya que "el modelo RDF no está adscrito a ningún formato de serialización particular" (Fernández, Martínez-Prieto, Arias Gallego, Gutiérrez, 2011). Estos formatos de serialización para anotar grafos RDF son lógicamente equivalentes y en el documento RDF 1.1 Primer [3] (W3C, 2014) se describe de forma introductoria los siguientes:

- Familia de lenguajes RDF Turtle (Terse RDF Triple Language)
    - *N-Triples* proporciona una manera sencilla para serializar grafos RDF mediante líneas de texto plano sin formato.
    - *Turtle* es una extensión de N-Triples. Introduce una serie de atajos sintácticos, como apoyo a los prefijos de espacio de nombres, listas y abreviaturas de los literales con tipificación de datos.
    - TriG es una extensión de Turtle que permite la especificación de múltiples gráficos en forma de un conjunto de datos RDF.
    - *N-Quads* es una extensión simple de N-Triples para el intercambio de conjuntos de datos RDF, permite añadir un cuarto elemento a una línea: el IRI del grafo de la tripleta descripta en esa línea.

- *JSON-LD* está diseñado en torno al concepto de un "contexto" para proporcionar mapeos adicionales desde JSON a un modelo RDF. El contexto vincula las propiedades del objeto en un documento JSON a los conceptos de una ontología. Con el fin de mapear la sintaxis JSON-LD para RDF, JSON-LD permite que los valores sean coaccionados a un tipo especificado o ser etiquetados con un lenguaje. Un contexto puede ser incrustado directamente en un documento JSON-LD o puesto en un archivo separado y referenciado desde diferentes documentos

- *RDFa* es una sintaxis RDF que puede utilizarse para incrustar datos RDF dentro de documentos HTML y XML, permitiendo, por ejemplo, que los motores de búsqueda agreguen estos datos al rastrear la web, enriqueciendo los resultados de búsqueda.

- *XML/RDF* especifica los triples RDF dentro del elemento XML rdf:RDF. Fue la única sintaxis de RDF cuando este modelo fue desarrollado en la década de 1990. En

[3] http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/

2001, se desarrolló N3 (Notation3) -antecedente de Turtle- y, poco a poco, fueron surgiendo los demás lenguajes que se han adoptado y estandarizado.

Modelado de datos

Las especificaciones para la definición de ontologías como OWL (Ontology Web Language) y otros vocabularios o sistemas de organización del conocimiento, como SKOS[4] (Simple Knowledge Organization System), permiten el modelado de los datos mediante la definición de clases y propiedades, así como de las relaciones entre estas.

Las ontologías permiten describir y representar conceptos dentro de un campo determinado de conocimiento, y definir las relaciones que se establecen entre ellos (Taylor, 2009). La W3C desarrolló un lenguaje para la definición de ontologías llamado OWL (Ontology Web Language[5]). OWL proporciona una semántica adicional que permite contribuir al objetivo de la interoperabilidad de los datos, tanto para su lectura por humanos como por agentes inteligentes (W3C, 2012).

RDFS (Resource Description Framework Schema)

Esquema RDF es una extensión de RDF para la definición de vocabularios.RDFS define clases de recursos en documentos RDF asignándoles propiedades.RDF Schema no especifica un vocabulario de propiedades descriptivas tales como “autor” o “título”, sino que especifica el mecanismo necesario para definir tales elementos,para definir las clases de recursos con los que se pueden utilizarse.

Lenguaje de consultas SPARQL (Simple Protocol and RDF Query Language)

Para alcanzar los objetivos de funcionar como un modelo para el enlazado de datos y la generación de relaciones lógicas por parte de los agentes inteligentes, RDF se construye también sobre la base de otras tecnologías asociadas, como lo son las ontologías para la generación de vocabularios controlados, y SPARQL[6] para la búsqueda y recuperación de información.
Como contrapartida a la generación de lenguajes para la descripción de la información, el W3C ha desarrollado también lenguajes para la interrogación de las bases de datos que contienen recursos descritos en RDF. El lenguaje de consulta SPARQL es una recomendación de la W3C desde el año 2008, y permite interrogar *data hubs* (concentrador de conjuntos de datos) o *data sets* (conjunto de datos sobre una temática específica) descritos en RDF, de forma tal de obtener información sobre los mismos. Para poder realizar la

[4] http://www.w3.org/TR/skos-reference/
[5] http://www.w3.org/TR/owl-features/
[6] http://www.w3.org/TR/rdf-sparql-query/

consulta SPARQL el *data hub* debe contener un SPARQL *endpoint*, y se debe conocer previamente la estructura del *data hub*, de forma de poder interrogar la base de datos de la forma adecuada (Prud'hommeaux, E. & Seaborne, A. (Eds.), 2008; W3C, 2013).

Intercambio de Información : RDF

RDF (Resource Description Framework / Marco de Descripción de Recursos:es un modelo estándar (W3C) para intercambio de datos en la web.Permite estandarizar la representación de los datos de forma significativa y comprensible para las máquinas, haciendo posible el intercambio de conjuntos de datos entre distintas plataformas y sistemas. Este modelo de información se basa en las relaciones lógicas entre las entidades (recursos).Fue desarrollado por World Wide Web Consortium (W3C) en el año 1997 (Taylor, 2009). Desde entonces se han desarrollado varias versiones hasta llegar a la RDF 1.1, la más actual del año 2014 que tiene el carácter de recomendación.

El estándar RDF permite describir e intercambiar recursos en la web mediante metadatos, no solo de documentos sino todo tipo de entidades. Puede utilizarse para diferentes aplicaciones, pero todas ellas coinciden en buscar mejorar la eficiencia en la recuperación de la información, facilitando la desambiguación de los recursos y permitiendo la búsqueda por parte de agentes inteligentes (robots, *crawlers*, etc.) capaces de interrogar las bases de datos en una enorme cantidad de datos disponibles (Senso, 2003).

RDF se compone de :

- una sintaxis abstracta, en la que se exponen los principales elementos de su modelo de datos;
- una sintaxis específica en la que se definen con mayor detalle estos elementos y su codificación en XML;
- una semántica que establece el significado de estos elementos o componentes.

La estructura de RDF

La estructura principal en la que se basa el modelo RDF es conocida con el nombre de tripleta. Una tripleta es una sentencia o declaración (*statement*) sobre un determinado recurso, que indica cierto conocimiento sobre el mismo. Un recurso puede ser una amplia cantidad de cosas, como un documento, un sitio web, un concepto o incluso un objeto de la vida real (Taylor, 2009).

La idea básica que subyace en la estructura de tripletas es que todos los recursos poseen propiedades específicas y que estas propiedades tendrán un valor particular para cada recurso.

Por lo tanto, todos los recursos pueden ser descritos en sentencias en las que se indiquen sus propiedades y los valores correspondientes (Beckett, 2004).

Según Cyganiak, R., Wood, D. & Lanthaler, M. (2014), entre otros, las tres partes que componen las tripletas son:

- Sujeto: recurso que se describe en la declaración.
- Predicado: propiedad o característica a la que se hace referencia en la declaración.
- Objeto: valor de la propiedad a la que se hace referencia en la declaración.

Por ejemplo, en la sentencia:

<El gaucho Martín Fierro> <tiene un autor que es> <José Hernández>

podemos observar los componentes:

- Sujeto: El gaucho Martín Fierro
- Predicado: tiene un autor que es
- Objeto: José Hernández

Esta sentencia resulta legible y comprensible para los humanos, pero no así para las computadoras y agentes inteligentes. Por tal motivo, el modelo RDF utiliza diferentes recursos para lograr que la sentencia completa pueda ser procesada por una máquina y, además, que cada una de estas partes de la sentencia pueda ser identificada de forma unívoca.

Las sentencias que están escritas de manera que puedan ser procesadas por un agente inteligente, emplean lenguajes como XML[7] (eXtensible Markup Language / Lenguaje de Marcas eXtensible) o JSON[8] (JavaScript Object Notation / Notación de Objetos de JavaScript), que permiten estructurar datos para su intercambio.

Además de las cuestiones asociadas a la sintaxis (notación) de las sentencias RDF, el estándar cuenta con una semántica que funciona como base del modelo y que se detalla en el trabajo de Hayes, P. & Patel-Schneider, P. (2014). De acuerdo a este documento todos los IRI que se utilicen en una declaración de RDF deben identificar siempre la misma entidad (recurso, persona, concepto, etc.)

Por otro lado, de acuerdo al documento, se da por sentado que toda la información declarada en una tripleta y las relaciones que ésta establece, son siempre verdaderas. Como resultado de esta semántica los agentes inteligentes pueden hacer inferencias lógicas en base a las

[7] http://www.w3.org/XML/
[8] http://www.json.org/json-es.html

declaraciones provistas en cada una de las tripletas (Cyganiak, R., Wood, D. & Lanthaler, M. 2014).

Representación de las sentencias mediante grafos

Todas las sentencias dentro del modelo RDF pueden representarse en un *grafo*. Esta estructura se compone de dos nodos (el sujeto y el objeto) y un arco (predicado) que los une mediante una relación. El grafo es siempre unidireccional, y se describe en el sentido Sujeto Predicado Objeto (Méndez Rodríguez, 2002).

Si se representa la sentencia utilizada en el ejemplo del apartado anterior en forma de grafo, obtendríamos la siguiente estructura:

Dentro de la estructura de grafo, cada uno de los tres componentes puede tener diversas formas de representación:

- Sujeto: puede ser un IRI o un nodo en blanco.
- Predicado: siempre es un IRI.
- Objeto: Puede ser un IRI, un nodo en blanco, o un literal.

A diferencia de los IRI, que funcionan como identificadores que permiten describir de manera unívoca un recurso, los literales se componen de cadenas de caracteres de representación léxica, números y fechas. En la estructura de grafo los IRI se diferencian de los literales en que los primeros son dibujados mediante figuras elípticas y los segundos mediante figuras rectangulares.

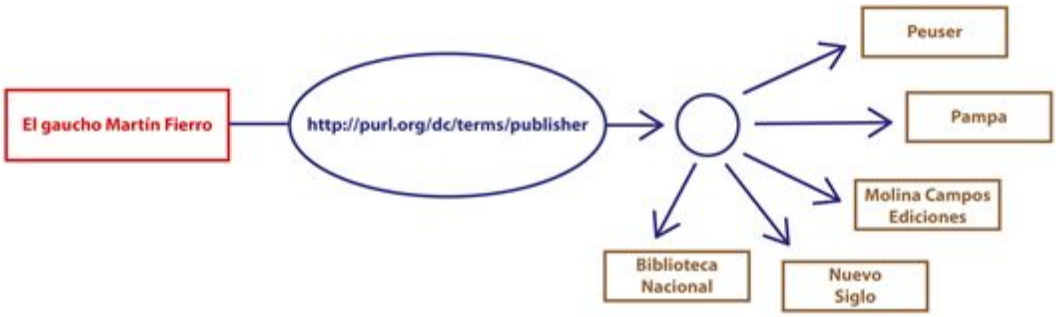
En el ejemplo anterior, para adecuar el gráfico a una representación de grafos RDF, deberíamos indicar la relación entre sujeto y objeto con un IRI, por ejemplo, tomándola de los elementos del esquema Dublin Core y podríamos cambiar el literal "José Hernández", por el IRI que identifica a este autor, por ejemplo, en el VIAF[9]:

Los literales deben estar siempre tipificados, es decir, acompañados del *datatype* (tipo de dato) utilizado. Por ejemplo, si a un literal numérico se le añade el *datatype integer* nos permite saber que consiste en un número entero, mientras que el *datatype date* nos permite saber que consiste en una fecha.

Por su parte los nodos en blanco son aquellos que no tienen un IRI que los identifique pero que pueden utilizarse para establecer relaciones entre un mismo sujeto y una variedad de objetos. De esta forma se pueden crear estructuras enlazadas que permiten expresar relaciones complejas (Taylor, 2009).

Retomando el ejemplo considerado anteriormente, pero relacionando la obra con sus múltiples editores, se podría utilizar un nodo en blanco para establecer la generalización :

[9] Virtual International Authority File (VIAF) https://viaf.org/



Etapa 3 : Herramientas de trabajo: FOSS de migración de datos

En la búsqueda de herramientas para la migración de datos, se comprobó que existe una gran cantidad que permite migrar registros en formato MARC21 o MARCXML a RDF. Para la selección de herramientas se plantearon los siguientes requisitos: que sean de código abierto, y de acceso libre y gratuito (FOSS-Free and Open Source Software), dado que estos sistemas ofrecen la posibilidad de replicar la experiencia sin el condicionamiento del pago de licencias. Sin embargo, se debe destacar que también existen conversores en línea y aplicaciones gratuitas pero no libres.

Se realizó una primera aproximación a la identificación de herramientas FOSS y se procedió a caracterizarlos. De este proceso surgirán tanto la selección de las herramientas como el primer borrador de la matriz de comparación. Asimismo, se están delineando los criterios para seleccionar la muestra de registros bibliográficos en formato MARC y la biblioteca de la cual se extraerán.

En este marco de trabajo se realizó una primera aproximación a las herramientas disponibles que cumplieran con estos requisitos. Se presentarán en esta ponencia cinco de ellas: *Easym2r, Catmandu, Xalan, Saxon y xsltproc.*

- *Easym2r* [10]es, según su desarrollador describe en la página del proyecto, un intento en lenguaje PHP de convertir datos en MARC a RDF. Solamente necesita los registros MARC, un archivo JSON-LD válido que muestre cómo debe quedar conformada la tripleta de RDF y tener instalado PHP 5.3 o superior en la PC donde correrá la aplicación.

[10] https://github.com/cKlee/easyM2R

- *Catmandu* [11]es un conjunto de módulos en lenguaje Perl destinada a facilitar la importación, almacenamiento, recuperación, exportación y transformación de registros de metadatos. Si bien este proyecto se propone evaluar únicamente las funciones de migración, Catmandu pretende convertirse en un ambiente de desarrollo para el diseño de servicios de bibliotecas digitales.

- *Xalan* [12] es una librería de Apache que implementa el lenguaje de transformación XSL "1.0 XML" y el Xpath. Puede utilizarse desde la línea de comandos o llamarse desde un programa externo. Anteriormente era distribuido de forma comercial (IBM LotusXSL), pero actualmente se distribuye bajo la licencia de Apache.

- *Saxon*[13] es un procesador de XSLT y XQuery creado por Michael Jay y desarrollado y mantenido por la compañía Saxonica. Existe una versión de código abierto y otra privativa. Presenta versiones para Java, JavaScript y .NET. Algunas versiones de Saxon son compatibles con XSLT 2.0 y hasta con 3.0, aunque la versión libre solamente permite utilizar XSLT 1.0 y 2.0, según la versión de Saxon que se instale. Es distribuido bajo la licencia de Mozilla y puede gestionarse directamente desde la línea de comandos de Linux.

- *Xsltproc*[14] es una herramienta que permite operar la librería en lenguaje C de libxslt utilizando la línea de comandos. Está basado en libxml2 y permite utilizar XSLT 1.0. Utiliza XML para el parseo de los datos y soporta XPath. Es distribuido bajo la licencia del Massachussets Institute of Technology (MIT).

[11] https://github.com/LibreCat/Catmandu
[12] https://xml.apache.org/xalan-j/
[13] http://saxon.sourceforge.net/
[14] http://xmlsoft.org/XSLT/xsltproc2.html

BIBLIOGRAFÍA

Beckett, D. (Ed.) (2004). RDF/XML Syntax Specification (Revised) : W3C Recommendation [en línea]. Disponible en: http://www.w3.org/TR/REC-rdf-syntax/ [Consulta 18/09/2015]

Caplan, P. (2003). *Metadata fundamentals for all librarians*. American Library Association.

Coyle, K. (2012). Semantic Web and linked data. *Library Technology Reports*,*48*(4), 10-14.
Colaboradores de Wikipedia. (2015) . Serialización [en línea]. Wikipedia, La enciclopedia libre. Disponible en <https://es.wikipedia.org/w/index.php?title=Serializaci%C3%B3n&oldid=83471626> [Consulta 18/09/2015]

Cyganiak, R., Wood, D. & Lanthaler, M. (Eds.) (2014). RDF 1.1 Concepts and Abstract Syntax : W3C Recommendation [en línea]. Disponible en: http://www.w3.org/TR/REC-rdf-syntax/ [Consulta 18/09/2015]

Daudinot Founier, I. (2006). Organización y recuperación de información en Internet: teoría de los metadatos. *Acimed*, *14*(5). Disponible en http://bvs.sld.cu/revistas/aci/vol14_5_06/aci06506.htm [Consulta: 18/09/2015]

Fernández, J. D., Martínez-Prieto, M. A., Arias Gallego, M., Gutiérrez, C. (2011). HDT·EndPoints: una Arquitectura Eficiente para la Web de Datos [en línea].
XVI Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2011), A Coruña, España, 5-7 de Septiembre de 2011.

Garzón Farinós, M. F. (2014). El registro de autoridades personales tras la aparición del web. PhD thesis, Universitat Politècnica de València (España). [Thesis]. Disponible en http://eprints.rclis.org/24571/ [Consulta: 18/09/2015]

Hayes, P. & Patel-Schneider, P (Eds.) (2014). RDF 1.1 Semantics : W3C Recommendation [en línea]. Disponible en: http://www.w3.org/TR/rdf11-mt/ [Consulta 18/09/2015]

Lamarca Lapuente, M. J. (2006). Hipertexto, el nuevo concepto de documento en la cultura de la imagen [en línea] (Tesis doctoral. Universidad Complutense de Madrid). URL: http://www. hipertexto. info [Consulta: 11/05/2009].

Méndez Rodríguez, E. (1999). RDF: un modelo de metadatos flexible para las bibliotecas digitales del próximo milenio. En: Jornades Catalanes de Documentació. Barcelona: Col.legi Oficial de Bibliotecaris Documentalistes de Catalunya, p. 487-498.

Méndez Rodríguez, E. (2002). Metadatos y recuperación de información: estándares, problemas y aplicabilidad en bibliotecas digitales. Gijón: Trea.

Nurseitov, N. Paulson, M. Reynolds, R. Izurieta, C. (2009). Comparison of JSON and XML Data Interchange Formats: A Case Study. ISCA 22nd International Conference on Computer Applications in Industry and Engineering, CAINE 2009, San Francisco, CA. Disponible en www.cs.montana.edu/izurieta/pubs/caine2009.pdf [Consulta 18/09/2015]

Pacheco, A., Ramírez, M., Guzmán, C., Cruz-Flores, R. (2013). Reproductores multimedia para la consulta de repositorios de recursos educativos abiertos desde dispositivos móviles. Disponible en http://www.researchgate.net/publication/263273638_Reproductores_Multimedia_para_la_Consulta_de_Repositorios_de_Recursos_Educativos_Abiertos_desde_Dispositivos_Mviles [Consulta 18/09/2015]

Prud'hommeaux, E. & Seaborne, A. (Eds.) (2008). SPARQL Query Language for RDF : W3C Recommendation [en línea]. Disponible en: http://www.w3.org/TR/rdf-sparql-query/ [Consulta 18/09/2015]

Quin, Liam. (2013). XML, JSON, XSLT and XQuery [en línea]. Disponible en: http://www.w3.org/blog/2013/09/xml-json-xslt-and-xquery/ [Consulta 18/09/2015]
Senso, J. A. (2003). Herramientas para trabajar con RDF. El profesional de la información, 12, (2), pp. 132-139.

Styles, Rob; Ayers, Danny; Shabir, Nadeem (2008). Semantic MARC, MARC21 and the Semantic Web [en línea]. Linked Data on the Web 2008 (LDOW2008). Disponible en: http://ceur-ws.org/Vol-369/paper02.pdf

Taylor, A. G. (2009). The organization of information. Westport, CO: Libraries Unlimited.

Voutssas Márquez, J., Barnard Amozorrutia, A. (2014). Glosario de Preservación Archivística Digital: Versión 4.0. México, D.F., Instituto de Investigaciones Bibliotecológicas y de la Información, UNAM (Tecnologías de la Información). Disponible en: http://goo.gl/HPBJhi [Consulta 18/09/2015]

W3C. (2012). OWL: Semantic Web Standards [en línea]. Disponible en http://www.w3.org/2001/sw/wiki/OWL. [Consulta 18/09/2015]

W3C. (2013). SPARQL 1.1 Overview: W3C Recommendation [en línea]. Disponible en http://www.w3.org/TR/rdf-sparql-query/ [Consulta 18/09/2015]

W3C. (2014). RDF 1.1 Primer [en lìnea]. Disponible en http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225 [Consulta 18/09/2015]