Diseño de un sistema para la generación de entradas de autoridad desde la perspectiva de los datos enlazados

Leandro Tabares-Martín[1], Amed Abel Leiva-Mederos[2], y Félix Oscar Fernández-Peña[3]

[1] Universidad de las Ciencias Informáticas
ltmartin@uci.cu
[2] Universidad Central "Marta Abreu" de las Villas
amed@uclv.edu.cu
[3] Instituto Superior Politécnico "José Antonio Hechavarría"
felix@ceis.cujae.edu.cu

Resumen El Control de Autoridades, en el caso de los nombres de los autores, se refiere a la generación de forma única de los mismos acorde a reglas previamente definidas. Esta es la parte más costosa del proceso de catalogación llevado a cabo por las instituciones bibliotecarias. Debido a la cantidad de información necesaria para llevar a cabo eficazmente esta actividad se hace necesario automatizar la recuperación de los datos requeridos, a la vez que se requiere una herramienta que realice la evaluación de las reglas para la generación de entradas de autoridad y brinde al usuario posibles resultados de forma que éste pueda seleccionar el más adecuado. En el presente trabajo se expone un diseño para la construcción de un sistema semi-automático con este objetivo, a la vez que se analiza una de las fuentes de datos seleccionadas para obtener la información.

Palabras Claves: Control de Autoridades, Datos Enlazados, Web Semántica

Abstract Authority Control, in the case of authors' name, refers to the generation of unique variants of them according to previously defined rules. This is the most expensive part in the cataloguing process achived by libraries institutions. Due to the high quantity of information needed to efficiently achive this activity, it is necessary to automatize the needed data retrieval, both a tool is required to evaluate the authority entries generation rules and give to the user the possible results in a way it can choose the most suitable option for the record. In this work is exposed the design for the construction of a semi-automatic system to achive this goal, both one of the selected data sources to retrieve the information is analized.

Keywords: Authority Control, Linked Data, Semantic Web

1 Introducción

El Control de Autoridades, en el caso de los nombres de los autores, se refiere a la generación de forma única de los mismos acorde a reglas previamente definidas. Esta ha sido un área de debate entre los bibliotecarios durante aproximadamente el último siglo y medio; numerosas son las opiniones sobre la forma en que deben registrarse los nombres de autores y los epígrafes de materia en los catálogos de las bibliotecas. Algunos han expuesto que esto no es necesario, que basta con registrar las entradas bibliográficas en el catálogo; otros siguen debatiendo el tema[1,2,3]. Lo cierto es que el volumen de información ha crecido exponencialmente en la última centuria, sobre todo con el advenimiento de las nuevas Tecnologías de la Información y las Comunicaciones (TIC), por esto el Control de Autoridades se hace vital a la hora de recuperar los registros. Abundantes son los ejemplos de redundancias en los cuales el nombre de los autores es registrado de diversas formas, dificultando en muchos casos la localización de sus publicaciones, la figura a continuación muestra un resultado de la búsqueda sobre José Martí en el Fichero Virtual Internacional de Autoridades (VIAF) que enlaza los datos de numerosas bibliotecas alrededor del mundo.


132 resultados encontrados para *Martí, José*

	Encabezamiento	Tipo
1	Martí, José, 1853-1895 Martí y Pérez, José Julián, 1853-1895 Martí, José	Autor personal
2	Berlanga, Luis G., 1921-2010 Berlanga, Luis G., 1921-.... Berlanga, Luis García García Berlanga, Luis (1921-2010) García Berlanga, Luis García-Berlanga Martí, Luis, 1921-2010	Autor personal
3	Martí i Bonet, Josep Maria Martí i Bonet, Josep M. (Josep Maria), 1937- Martí Bonet, José M. Martí i Bonet, Josep M. Martí i Bonet, Joseph Maria Martí Bonet, J. María	Autor personal
4	González Casanova, José Antonio González Casanova, J. A. (Josep Antoni), 1935- González Casanova, José Antonio, 1935-.... González Casanova, J. A. 1935-	Autor personal

Figura 1. Búsqueda sobre José Martí en VIAF

Desde la década de 1970 las instituciones bibliotecarias han declarado que el Control de Autoridades es la parte más costosa del proceso de catalogación y todavía se buscan formas de automatizarlo para simplificar sus costos[4,5,1,3]. Un paso gigante en esta dirección lo constituye el compartir los ficheros de autoridades entre las bibliotecas. Ejemplo de esto es el Proyecto Cooperativo de Nombres de Autoridades (NACO, por sus siglas en Inglés)[5] que involucró a la Biblioteca

del Congreso de los EE.UU y otros socios. Las tecnologías actuales ofrecen nuevas oportunidades para enlazar esos ficheros de autoridades, mejorarlos y crear nuevos servicios para los usuarios[2].

Producto a la alta complejidad existente en el proceso de generación de entradas de autoridad, se hace necesario el desarrollo de un sistema informático que permita recopilar la información necesaria para posteriormente evaluar las Reglas de Catalogación Angloamericanas (RCAA)[6] y ofrecer al usuario variantes acertadas para la generación de una entrada de autoridad, facilitándole la selección de la más adecuada para el registro que está generando. Para la correcta generación de las entradas de autoridad es necesario contar con datos como el nombre y apellidos del autor, títulos de nobleza, estado civil, entre otros. Esta información se encuentra dispersa geográficamente, a la vez que no está descrita de forma uniforme en los diferentes conjuntos de datos.

Las Reglas de Catalogación Angloamericanas, en su capítulo dos, establecen cómo deben conformarse los "Encabezamientos de Personas"[6]; las mismas recogen una gran cantidad de posibles variantes, sin embargo, debido a la dificultad en la localización de la información para evaluarlas o al desconocimiento de la totalidad de las Reglas, se generan de forma cotidiana entradas de autoridad ambiguas.

Con la finalidad de facilitar la correcta generación de entradas de autoridad se persigue como objetivo de este trabajo el diseño de un Sistema que aproveche las posibilidades brindadas por las tecnologías de la Web Semántica, para contribuir al Control de Autoridades en Sistemas Integrados de Gestión Bibliotecaria (SIGB). A su vez se analiza brevemente la estructura de los datos aportados por la Biblioteca Nacional de España (BNE), la cual brinda valiosos recursos de información en forma de Datos Enlazados que serán utilizados como una de las fuentes de datos para el Sistema.

En el presente trabajo se relacionan algunos trabajos realizados respecto a los elementos bases de la Web Semántica y al Control de Autoridades desde esta perspectiva, se analizan los datos aportados por la Biblioteca Nacional de España, se presenta el diseño concebido para el desarrollo del Sistema y se plantean los nuevos retos que serán afrontados. Se considera que la utilización de un Sistema como el expuesto en este trabajo contribuirá a lograr una mayor normalización en esta área de los procesos bibliotecarios.

2 Trabajos relacionados

A causa de la cantidad de información necesaria para la correcta generación de entradas de autoridad el proceso de recuperación de la misma se hace complejo. Las tecnologías de la Web Semántica [7] permiten describir semánticamente los contenidos, de forma que sea posible delegar el trabajo de recuperación en aplicaciones informáticas desarrolladas para este fin y estas tengan un alto nivel de precisión con respecto a la información recuperada.

Las ontologías[8] permiten representar de manera formal los datos, de esta forma las aplicaciones pueden recuperar de forma más precisa los datos nece-

sarios para su funcionamiento. Estas posibilidades fueron aprovechadas por la Federación Internacional de Asociaciones Bibliotecarias (IFLA, por sus siglas en Inglés) construyendo ontologías que describiesen los datos de los registros bibliográficos utilizados por las bibliotecas. Producto de este esfuerzo surgieron ontologías como Requisitos Funcionales para Registros Bibliográficos (FRBR, por sus siglas en Inglés)[9], Requisitos Funcionales para Datos de Autoridad (FRAD, por sus siglas en Inglés)[10] entre otras.

El Esquema para la Descripción de Metadatos de Autoridad en RDF (MADS/RDF, por sus siglas en Inglés)[11] es un modelo de datos para datos de autoridad y vocabularios utilizados en las Ciencias de la Información y Bibliotecología (CIB) creado por la Biblioteca del Congreso de los EE.UU. MADS/RDF es un Sistema de Organización del Conocimiento basado en el Modelo Simple de Organización del Conocimiento (SKOS, por sus siglas en Inglés) [12] adaptado a las necesidades específicas de las CIB.

La Biblioteca del Congreso de los Estados Unidos desarrolló un marco de trabajo que denominó BIBFRAME[13]. Este Marco de Trabajo pretende representar la información bibliográfica en forma de Datos Enlazados Abiertos, utilizando para esto vocabularios definidos por la IFLA como FRBR[9]. Esta forma de estructuración de los datos facilita el descubrimiento de los mismos, a la vez que permite su re-utilización por sistemas como el propuesto y vínculo con otros datos que aportan mayor información a la hora de describir el registro bibliográfico.

El Fichero Virtual Internacional de Autoridades es un servicio internacional diseñado para permitir el acceso a los mayores ficheros de autoridad del mundo, fue creado con el objetivo de posibilitar el acceso desde la Web a los registros de autoridades de numerosas bibliotecas alrededor del mundo utilizando las tecnologías base de la Web Semántica. Para brindar sus servicios VIAF enlaza los ficheros de autoridades de varias bibliotecas nacionales y agrupa sus contenidos bajo una entrada de autoridad común[14], razón por la que debe tenerse en cuenta e investigar formas de consumir los datos que aporta.

"Amigo de un amigo" (FOAF, por sus siglas en Inglés)[15] es una iniciativa que persigue el objetivo de crear una Web de páginas "entendibles" por las computadoras. Encontrar formas de integrar este tipo de iniciativas con los mecanismos para el Control de Autoridades en las bibliotecas, puede contribuir a introducir los catálogos de los SIGB entre las diferentes herramientas disponibles en la Web. Adicionalmente, la disponibilidad de las entradas de autoridad bibliotecarias en una forma más orientada a la Web, tiene el potencial para contribuir positivamente en la organización del amplio espectro de información disponible en la Web[16]. La existencia de FOAF fue contemplada en la creación de BIBFRAME[13] por las posibilidades que brinda al permitir relacionar personas registradas dentro de un registro catalográfico con otros recursos disponibles en la Web, facilitando el descubrimiento de nueva información en ambas direcciones.

AUTHORIS es una herramienta desarrollada en la Universidad de Granada que aspira a facilitar el procesamiento de datos de autoridad de una manera

estandarizada basándose en los principios de los Datos Enlazados[17], centrada en el uso de reglas de aprendizaje automático y las posibilidades de los Datos Enlazados para operar registros de diversas organizaciones. Está basada en el sistema para la administración de contenidos Drupal y aprovecha sus facilidades para publicar datos en el Marco de Trabajo para la Descripción de Recursos (RDF, por sus siglas en Inglés)[18]. Esta herramienta constituye un paso de avance con el fin de automatizar la generación de entradas de autoridad, sin embargo, se trata de una herramienta de consulta que no posibilita la interacción directa con un SIGB, obligando al usuario a localizar la entrada de autoridad correspondiente y posteriormente trasladarla hacia su registro catalográfico.

3 Conjunto de datos de la Biblioteca Nacional de España

La Biblioteca Nacional de España, de conjunto con el "Ontology Engineering Group" de la Escuela Técnica Superior de Ingeniería Informática de la Universidad Politécnica de Madrid, exportó sus Registros de Autoridades y Registros Bibliográficos a formato de Datos Enlazados[19] y los ofrece para su descarga gratuita a la vez que brinda un punto de acceso para realizar consultas. Los Identificadores de Recursos Internacionalizados (IRIs, por sus siglas en Inglés) de estos grafos se enuncian en la Tabla 1.

Cuadro 1. IRIs de los grafos de autoridades y registros bibliográficos de la Biblioteca Nacional de España

Nombre del grafo	*IRI*
Registros de Autoridades	http://datos.bne.es/graph/dataset/autority
Registros Bibliográficos	http://datos.bne.es/graph/dataset/bibliographic

Para la representación de los datos ofrecidos por la BNE se utilizaron ontologías definidas por la Federación Internacional de Asociaciones Bibliotecarias e Instituciones. En este conjunto de datos se utilizan ontologías como FRBR[9], FRAD[10] y los Estándares Internacionales para Descripción Bibliográfica (ISBD, por sus siglas en Inglés)[20].

4 Diseño del Sistema para la generación de entradas de autoridad

Para desarrollar el Sistema para la Generación de Entradas de Autoridad se precisa implementar las reglas para la generación de Encabezamientos de Personas, de forma que las mismas puedan ser evaluadas computacionalmente. Estas reglas evaluarán la información obtenida por un conjunto de recuperadores especializados que serán implementados con el propósito de realizar consultas acordes

a las características de los vocabularios utilizados para describir la información almacenada en cada una de las fuentes de datos.

Con el objetivo de explotar la posibilidad de realizar consultas optimizadas se utiliza para almacenar los grafos de tripletas RDF el "Virtuoso Universal Server", ya que este es el mismo utilizado para la publicación de sus datos por parte de la Biblioteca Nacional de España. La arquitectura de este servidor permite la persistencia de datos en diferentes formas como por ejemplo: Relacional, RDF, XML, texto plano y Datos Enlazados. En escenarios de integración de datos en que se necesita recuperar información desde fuentes de datos remotas "Virtuoso Universal Server" provee una combinación de vistas SQL (Lenguaje de Consultas Estructurado) y SPARQL[21] usando uniones federadas que simulan el acceso a una base de datos única, a la vez que un optimizador de consultas interviene en el momento de encuestar las bases de datos teniendo en cuenta si son locales o remotas[22].

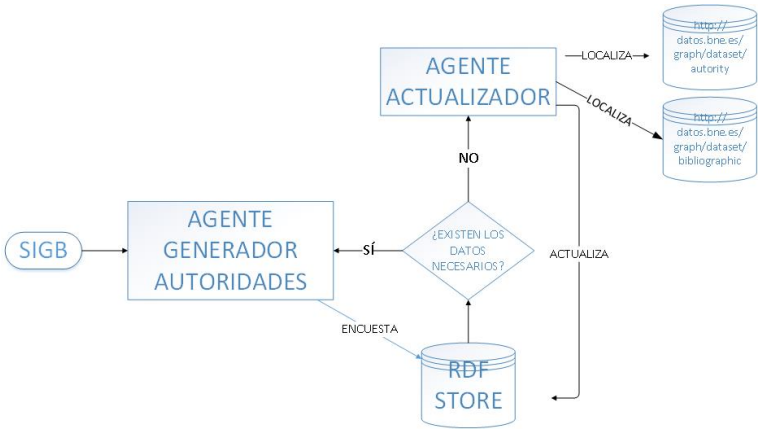

Figura 2. Diseño del Sistema para la Generación de Entradas de Autoridad

El Sistema se compone de dos agentes de software:

1. Agente Actualizador: Encargado del enriquecimiento de los datos almacenados localmente a partir de los elementos localizados en línea que son necesarios para la evaluación de las reglas por parte del Agente Generador de Autoridades. Estos datos serán almacenados en un grafo aparte que utilizará para la descripción de la información vocabularios estandarizados por la Federación Internacional de Asociaciones Bibliotecarias (IFLA, por sus siglas en Inglés) como FRBR, FRAD e ISBD. Este agente encuesta el grafo de registros bibliográficos de la BNE a partir de los datos provistos por el SIGB y obtiene el identificador del registro de autoridad del autor en el grafo de autoridades, luego encuesta el grafo de autoridades y obtiene la información

necesaria para que el Agente Generador de Autoridades evalúe las reglas y genere la entrada de autoridad correspondiente.
2. Agente Generador de Autoridades: Este Agente es accesible vía un servicio web y permite la integración de diferentes Sistemas Integrados de Gestión Bibliotecarios (SIGB) con el Sistema. Es el encargado de evaluar las reglas establecidas por las RCAA2 y generar las entradas de autoridad para el SIGB que las requiera.

Los agentes han sido implementados utilizando tecnología Java en su versión 1.7 funcionando sobre la máquina virtual OpenJDK. GNU Prolog para Java[23] es una implementación del lenguaje Prolog utilizable como una librería del lenguaje Java, esta fue empleada para evaluar las reglas dada las capacidades de inferencia de este lenguaje. Para desacoplar el Sistema de un SIGB en específico fue utilizada la tecnología provista por Apache CXF[24], el cual es un marco de trabajo de código abierto para servicios que provee una variedad de protocolos como SOAP[25], XML/HTTP[26] y RESTful HTTP[27].

Para la recuperación de la información a partir del conjunto de datos ofrecidos por la BNE se implementó un "crawler" que permite extraer y transformar los datos acorde a las necesidades para su posterior procesamiento. El mismo utiliza el API ARQ que provee Jena[28] para la generación de las consultas SPARQL desde Java. El Sistema ha sido integrado con el SIGB ABCD que se desarrolla como parte de un proyecto de Investigación y Desarrollo en la Universidad de las Ciencias Informáticas en Cuba.

Al estar desacoplado el Sistema de un SIGB en particular, permite que los sistemas que deseen utilizar sus servicios solamente tengan que consumir un servicio web. Si el SIGB que lo encuesta es capaz de enviarle el título de la obra que se está catalogando, el año de publicación y su Número Internacionl Estándar de Libro (ISBN, por sus siglas en Inglés), el Sistema recupera los datos del autor de dicha obra y genera la entrada de autoridad correspondiente.

5 Conclusiones y trabajo futuro

Con el objetivo de aprovechar las posibilidades ofrecidas por los Datos Enlazados, instituciones como la Biblioteca Nacional de España han exportado sus registros de autoridades y bibliográficos conforme a esta estructura. El contar con fuentes de datos estructuradas como Datos Enlazados posibilita que los nuevos SIGB que se desarrollen aprovechen las bondades que ofrece. A su vez los futuros SIGB deberían contemplar la posibilidad de utilizar los conjuntos de datos puestos libremente a disposición de la comunidad internacional y compartir los suyos propios en forma de Datos Enlazados Abiertos, así se podrá lograr una mayor normalización en este campo y se dará un paso de avance importante en el desarrollo de una "catalogación compartida". Se pretende posteriormente analizar otras fuentes de datos compartidas en forma de LOD, como lo son los Registros de Autoridad de la Biblioteca del Congreso de los EE.UU.

Referencias

1. Warnner, J.W., Brown, E.W.: Automated name authority control. In: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries. (2001) 21–22
2. Tillett, B.B.: Authority control: State of the art and new perspectives. Cataloging & classification quarterly **38**(3-4) (2004) 23–41
3. Gorman, M.: Authority control in the context of bibliographic control in the electronic environment. Cataloging & Classification Quarterly **38**(3-4) (2004) 11–22
4. JULAC-Project: HKCAN (1999)
5. Fenly, J.G., Irvine, S.D.: The name authority co-op (NACO) project at the library of congress: Present and future. Cataloging & classification quarterly **7**(2) (1986) 7–18
6. Gorman, M.: The Concise AACR2: Being a Rewritten and Simplified Version of Anglo-American Cataloguing Rules. American Library Association (1981)
7. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific american **284**(5) (2001) 28–37
8. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge acquisition **5**(2) (1993) 199–220
9. Associations, I.F.o.L., Committee, I.S.o.C.S., Records, I.S.G.o.t.F.R.f.B.: Functional requirements for bibliographic records: final report. Volume 19. KG Saur Verlag Gmbh & Company (1998)
10. Patton, G.E.: Functional requirements for authority data: A conceptual model. Volume 34. Walter de Gruyter (2009)
11. LOC: MADS/RDF primer (May 2012)
12. Miles, A., Matthews, B., Wilson, M., Brickley, D.: SKOS core: simple knowledge organisation for the web. In: International Conference on Dublin Core and Metadata Applications. (2005) pp–3
13. LOC: Bibliographic framework as a web of data: Linked data model and supporting services (November 2012)
14. OCLC: VIAF (February 2013) The Virtual International Authority File.
15. Brickley, D., Miller, L.: The Friend of a Friend (FOAF) project. FOAF Project (2000)
16. Harper, C.A., Tillett, B.B.: Library of congress controlled vocabularies and their application to the semantic web. Cataloging & classification quarterly **43**(3-4) (2007) 47–68
17. Leiva-Mederos, A., Senso, J.A., Domínguez-Velasco, S., Hípola, P.: AUTHORIS: a tool for authority control in the semantic web. Library Hi Tech **31**(3) (2013) 536–553
18. Lassila, O., Swick, R.R.: Resource description framework (RDF) model and syntax specification. W3C recommendation (1999)
19. Vila-Suero, D., Villazón-Terrazas, B., Gómez-Pérez, A.: datos. bne. es: A library linked dataset. Semantic Web **4**(3) (2013) 307–313
20. IFLA: International standard bibliographic description | IFLA (2011)
21. Consortium, W.W.W.: SPARQL 1.1 overview (March 2013)
22. Williams, P.: An overview of virtuoso universal server (February 2013)
23. Project, G.: GNU prolog for java (2010)
24. Foundation, A.S.: Apache CXF (2013)
25. Consortium, W.W.W.: SOAP specifications (April 2007)
26. Consortium, W.W.W.: XMLHttpRequest level 1 (January 2014)
27. Roth, G.: RESTful HTTP in practice (August 2009)
28. Foundation, A.S.: Apache jena - ARQ - a SPARQL processor for jena (2013)