F. 9. Ontologías de control de autoridades en el ámbito de los datos abiertos enlazados

Juan-Antonio Pastor-Sánchez

10 noviembre 2012

Pastor-Sánchez, Juan-Antonio (2013). "Ontologías de control de autoridades en el ámbito de los datos abiertos enlazados". *Anuario ThinkEPI*, v. 7, pp. 184-188.

Resumen: A partir del papel que los lenguajes documentales están desempeñando en el ámbito de los datos abiertos enlazados, se realiza un análisis de dos ontologías cuyo objetivo es la publicación en este entorno de ficheros de control de autoridades: *Mads/RDF* y *GND*. Finalmente se reflexiona sobre la necesidad de contemplar la realidad de los datos abiertos enlazados desde una perspectiva más integradora y global, en donde además de los conjuntos de datos, también estén interconectados los vocabularios controlados y los modelos de descripción.

Palabras clave: Datos abiertos enlazados, Ontologías, Autoridades, *Mads*, *GND*, *Skos*.

Title: Authority control ontologies in the field of linked open data

Abstract: Based on the approach of the role that documentary languages develop in the linked open data environment, two ontologies for publishing authority control files are briefly analysed: *Mads/RDF* and the *GND* ontology. We consider the need to contemplate the linked open data reality in a more inclusive and global perspective, in which not only datasets are interconnected, but also controlled vocabularies and description models.

Keywords: Linked open data, Ontologies, Authorities, *Mads*, *GND*, *Skos*.

Introducción

La aparición, aceptación y despliegue de una nueva tecnología conlleva nuevas oportunidades de evolución para aplicaciones que en un principio parecían haberse quedado obsoletas o relegadas a ámbitos muy reducidos. En un mundo dominado por los motores de búsqueda web, los vocabularios controlados fueron relegados a un segundo e incluso tercer plano. Su aplicación a la indización y clasificación automatizadas ha resultado en cierto sentido infructuosa y el surgimiento de los buscadores ofreció en su momento una alternativa eficiente (aunque poco precisa) y sencilla.

El desarrollo de modelos de metadatos y su despliegue mediante estándares de la web semántica abre nuevas perspectivas de uso de los lenguajes documentales, vocabularios controlados y sistemas de organización del conocimiento en general. Surge la posibilidad de dotar de valor añadido a los vocabularios controlados mediante su publicación y su uso abiertos. Un determinado tesauro, fichero de autoridades o lista de encabezamientos de materia tiene más valor si se publica bajo el paradigma de los datos abiertos enlazados puesto que se facilita su interoperabilidad y, en consecuencia, su reutilización y aplicación en cualquier otro ámbito.

"Los ficheros de control de autoridades son una herramienta bastante desaprovechada"

Dentro de la diversidad de los vocabularios controlados, los ficheros de control de autoridades tal vez sean los que estén más ligados al proceso de catalogación, puesto que su génesis se produce dentro de este proceso. Si bien su creación y estructuración no parten de los principios que rigen la organización del conocimiento, su incardinación en esta disciplina es indiscutible. Las clasificaciones, tesauros y encabezamientos de materia presuponen una estructura conceptual interna que, aunque a veces hayan utilizado corpus documentales para su elaboración, precede a la indización. Las autoridades se identifican, extraen y normalizan durante la descripción bibliográfica, cuyo control se centraliza en un fichero o registro de control de autoridades de carácter interno o externo.

Las autoridades abarcan muchos más aspectos que la descripción temática, como los referidos a personas, ubicaciones geográficas, organizaciones o eventos. Son una herramienta más "completa" que un tesauro, aunque también mucho más heterogénea. Su función inicial no es la organización del conocimiento subyacente a la colección bibliográfica a la que hace referencia, sino facilitar la localización y consulta de un registro específico en un catálogo a través de la unificación y desambigüación de distintas formas de un mismo punto de acceso.

La emergencia de este tipo de herramientas podría situarse en los sistemas integrados de gestión bibliotecaria, especialmente en su aplicación dentro de los opacs online (**Herrero**, 1999). Su aplicación se ha orientado a la mejora del proceso de recuperación de información en estos entornos.

Sin embargo, tal y como indica **Rodríguez-Yunta** (2012), los ficheros de control de autoridades son una herramienta bastante desaprovechada. Esto puede deberse en parte a que son poco visibles o tienen poca incidencia durante el proceso de búsqueda por parte del usuario. Como apunta el mismo autor: al final todo el software de búsqueda desea adoptar el aspecto de un motor, como *Google*, cuando tal vez se precise otro tipo de funcionalidad que oriente al usuario durante el proceso de consulta del catálogo utilizando precisamente el control de autoridades.

Desde el punto de vista de la reutilización de este recurso, los ficheros de autoridades constituyen el producto de una labor de descripción documental que va más allá de la descripción temática. La extracción de puntos de acceso y su concordancia en un registro de autoridades podría ser explotado por parte de otras instituciones mediante su publicación como datos abiertos enlazados (**Peset**; **Ferrer-Sapena**; **Subirats-Coll**, 2011).

De hecho, ésta es la base sobre la que se han desarrollado algunos conjuntos de datos como *Viaf*, *SWD/GND*, *Geonames*, *LC/NAF*, *Ndlsh* o *Rameau* entre otros. A este respecto es recomendable la consulta del documento elaborado por el *Grupo Incubadora de Datos Enlazados Bibliotecarios* del *W3C* (**Isaac** *et al.*, 2011).

La reutilización de esta información pasa necesariamente por la adopción de una estructura que permita su publicación como *linked open data* y son varias las propuestas que actualmente en forma de ontologías y esquemas *RDF* se perfilan como las opciones más adecuadas para este fin. En esta nota se analizan brevemente dos de ellas: *Mads* y *GND*.

Mads/RDF

El origen de *Mads/RDF* se encuentra en un esquema xml con el mismo nombre genérico. Es compatible con *Marc21* y conforma el complemento de *Mods*[1] para la descripción de autoridades. Esta ontología se ha realizado por la *Library of Congress de Estados Unidos* y actualmente define 58 clases y 72 propiedades[2].

La propuesta de *Mads/RDF* es mucho más compleja de la que nos puede ofrecer una ontología de propósito más general como *skos*[3]. Destaca su capacidad para agrupar, mediante un mecanismo de precoordinación, diferentes entradas de autoridades en un tipo más complejo.

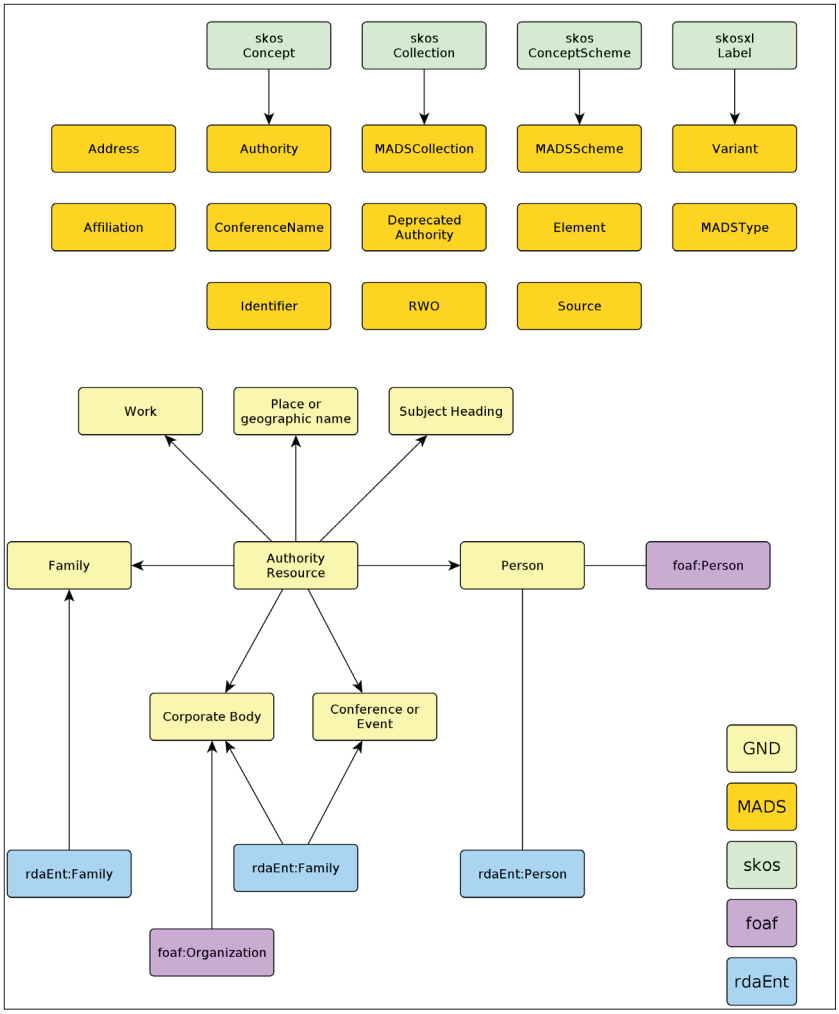

Figura 1. Estructura de clases principales de *Mads/RDF* y la ontología *GND* y su mapeado con otros esquemas

Por ejemplo: a partir de conceptos de distinto tipo (geográficos, temáticos, cronológicos) como "España", "Guerra civil", "Historia" y "1936-1937", es capaz de definir mediante la *clasemadsrdf:ComplexSubject* una autoridad compuesta bajo la etiqueta de "España – Historia – Guerra civil – 1936-1937". Este mecanismo también se aplica mediante clases específicas para representar relaciones jerárquicas entre conceptos geográficos (*madsrdf:HieraquicalGeographic*) y parejas de conceptos de tipo nombre-título (*madsrdf:NameTitle*).

Otro aspecto interesante de *Mads/RDF* es la identificación de autoridades en desuso, y un modo muy sencillo para el control de las distintas formas de una misma autoridad con los correspondientes reenvíos a la entrada normalizada de la autoridad, algo imprescindible para su aplicación en entornos de búsqueda online.

Mads/RDF está parcialmente mapeado con *skos*. Esto significa que para algunos de los elementos del vocabulario de *Mads/RDF* se han definido las oportunas equivalencias con los elementos de *skos*[3]. Más concretamente, la clase *madsrdf:Authority* se ha definido como equivalente de la clase *skos:Concept*. Otro ejemplo: la propiedad *madsrdf:authoritativeLabel* está mapeada con *skos:prefLabel*. No obstante hay que insistir en la parcialidad de dicho mapeado, ya que muchas de las clases y propiedades de *Mads/RDF* no tienen una equivalencia directa con *skos*.

Ontología *GND*

Gemeinsame Norm Datei es una propuesta de la *Biblioteca Nacional Alemana*. En principio se creó para la representación y publicación como datos abiertos enlazados de los elementos de *GND*, un fichero que integra autoridades de entidades, nombres personales, encabezamientos de materia y títulos. Cuenta con 50 clases, 162 propiedades para la definición de relaciones entre objetos y 56 para la definición de atributos mediante tipos de datos específicos[4]. Se trata de una ontología muy detallada, algunos de cuyos elementos están mapeados con elementos de vocabularios *RDA* y con *foaf*[5].

Cabe destacar el alto nivel de granularidad de esta ontología (desde el punto de vista de la descripción conceptual), así como la disponibilidad de clases y propiedades centradas en la descripción de conferencias y eventos. Un ejemplo del nivel de detalle que ofrece puede encontrarse en la definición de diferentes subpropiedades derivadas a partir de la propiedad general *gnd:associatedPlace*; en este sentido podemos encontrar subpropiedades para la descripción de lugares de nacimiento, muerte, exilio, actividad económica, creación, custodia, etc. También ofrece distintos tipos de autoría como pintor, poeta, inventor, escritor, etc., para definir la relación de una persona o institución en el proceso de creación de una obra.

La ontología *GND* es totalmente compatible con *Marc21*. De hecho dispone de la propiedad de anotación *gnd:marc21equivalent*, a través de la cual se han definido las equivalencias de cada elemento de la ontología con los campos correspondientes de dicho formato.

Pero ¿cuál es mejor?

No hay una respuesta sencilla. Ambas propuestas tienen planteamientos muy diferentes. *Mads/RDF* se adapta en mayor medida a una concepción más tradicional de los ficheros de autoridades, al tiempo que añade interesantes propiedades de agrupación y una correspondencia parcial con *skos*. Esto último permite la alineación entre registros de autoridades representados con *Mads/RDF* y otros conjuntos de datos con vocabularios controlados que usen *skos*.

"La reutilización de los ficheros de control de autoridades pasa por adoptar una estructura que permita su publicación como *linked open data*"

Por su parte *GND* va más allá, constituyendo un considerable esfuerzo para la definición de dominios de conocimiento a partir de un instrumento (los ficheros de autoridades) que en principio no fue creado para ello. Este aspecto, junto con la meticulosa definición de equivalencias con *Marc21* y con vocabularios *Mads/RDF*, ofrece una interesante combinación que hace de esta ontología un puente entre una aplicación más clásica y el futuro próximo que ya está tocando a nuestra puerta. Pese a ello sería recomendable definir las relaciones de mapeado pertinentes con *skos* para facilitar la interoperabilidad de los conjuntos de datos que opten por aplicar la ontología.

La decisión de usar una u otra opción dependerá de las expectativas de puesta en valor que se tenga sobre un fichero de control de autoridades. La aplicación de *Mads/RDF* puede resultar algo más inmediata y por supuesto imprescindible en entornos en donde se utilice *Mods*. La explotación de estos conjuntos de datos se orienta más hacia procesos de búsqueda y recuperación de información online. Con *GND* puede aplicarse un amplio repertorio de clases y propiedades para representar dominios de conocimiento que per-

mitirían la realización de procesos de inferencia que podrían ser utilizados no solamente en tareas de búsqueda de información, sino también en el descubrimiento de información. No obstante su principal handicap es el gran número de elementos, que puede hacer que su uso resulte bastante complejo.

A todo lo anterior hay que añadir la existencia de *skos* como un modelo a partir del cual se pueden derivar soluciones para el caso que nos ocupa. En tal sentido, se está trabajando para adaptar *skos* a *ISO 25964*, debiendo considerarse que el modelo de datos propuesto por esta nueva norma de tesauros podría aplicarse para el control de autoridades[6].

Necesidad de una visión más amplia sobre datos abiertos enlazados

La publicación de los ficheros de autoridades como datos abiertos enlazados muestra un nuevo panorama en el que los ficheros de autoridades evolucionan desde una herramienta de control terminológico hacia otra más centrada en la interconexión de conjuntos de datos. Tal vez la puesta en valor de estos vocabularios sea el primer paso hacia la publicación de los propios catálogos. Esto permitiría cambiar la situación actual señalada por el informe final del *Grupo Incubadora* sobre datos enlazados bibliotecarios (**Baker**, 2011): existen muchos más conjuntos de datos disponibles con vocabularios controlados que de catálogos y datos bibliográficos en general.

Por otro lado una misma necesidad (publicación de autoridades como datos abiertos) puede dar lugar al desarrollo de diferentes soluciones (*Mads/RDF* y *GND*). Cabe llegar a la conclusión de que nos dirigimos hacia una dispersión de modelos descriptivos, esquemas de metadatos y ontologías. Esta diversidad de alternativas para realizar una misma tarea puede crear confusión con respecto a la aplicación de una u otra opción.

Esta situación se puede evitar con una nueva aproximación a los datos abiertos enlazados que supere su concepción como meros conjuntos de datos creados y mantenidos por separado, entre los que se establecen enlaces para enriquecer los posteriores procesos de búsqueda y descubrimiento de información. Hay que adoptar una visión más integral y amplia de *linked open data* en la que además, y como norma general, se definan relaciones entre diferentes vocabularios controlados, así como entre ontologías o esquemas de metadatos. Sería necesario por tanto definir las correspondencias entre los elementos de *Mads/RDF* con la ontología *GND* y ésta a su vez con *skos*. Incluso habría que plantear la definición de equivalencias con otros esquemas y ontologías más generales, como por ejemplo las que dan soporte a *Geodata* o *DBpedia*.

Se trata de una filosofía subyacente de mapeado o alineación que sentaría las bases de la evolución futura de los datos abiertos enlazados y la elaboración de nuevos modelos descriptivos. De no llevarse a cabo es posible que los datos de bibliotecas, archivos y museos queden aislados de otros conjuntos de datos más globales que ya cuentan con un mayor grado de interconexión y por tanto de mayor utilidad y aplicación en la difusión y generación de conocimiento.

"Hay que adoptar una visión más integral de *linked open data*"

Notas

1. *Mods* (*metadata object description schema*) es un esquema xml para la descripción de recursos; y *Mads* (*metadata authority description schema*) es igualmente un esquema xml para la codificación de descripciones de autoridades. Ambos están integrados aunque pueden utilizarse por separado. Más información en:
http://www.loc.gov/standards/mods/design-principles-mods-mads.html

2. Para más información se pueden consultar la guía de uso de *Mads/RDF* y la referencia completa de la ontología:
http://www.loc.gov/standards/mads/rdf
http://www.loc.gov/standards/mads/rdf/v1.html

3. *Skos* (*simple knowledge organization system*) es una ontología para la publicación de sistemas de organización del conocimiento. Más información en:
http://skos.um.es/TR/skos-primer

4. La referencia completa de la ontología se encuentra en:
http://d-nb.info/standards/elementset/gnd

5. Es posible consultar más información sobre vocabularios *RDA* y *foaf* en:
http://rdvocab.info
http://xmlns.com/foaf/spec

6. El grupo de trabajo *ISO TC46/SC9/WG8*, responsable de la creación de la norma *ISO 25964*, está trabajando conjuntamente con **Antoine Isaac**, coeditor de la recomendación *skos* y el mapeado entre *Mads-skos*. Más información en:
http://www.niso.org/schemas/iso25964/correspondencesSKOS

Referencias bibliográficas

Baker, Thomas *et al*. (2011). *Library linked data incubator group final Report. W3C Incubator Group Report 25 October 2011*.
http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025

Herrero-Pascual, Cristina (1999). "El control de autoridades". *Anales de documentación*, n. 1, pp. 121-136. *http://revistas.um.es/analesdoc/article/viewFile/2621/2601*

Isaac, Antoine; **Waites, William**; **Young, Jeff**; **Zeng, Marcia** (2011). *Library Linked Data Incubator Group: datasets, value vocabularies, and metadata element sets. W3C Incubator Group Report, 25 de octubre. http://www.w3.org/2005/Incubator/lld/XGR-lld-vocab-dataset-20111025*

Peset, Fernanda; **Ferrer-Sapena, Antonia**; **Subirats-Coll, Imma** (2011). "*Open data* y *linked open data*: su impacto en el área de bibliotecas y documentación". *El profesional de la información*, marzo-abril, v. 20, n. 2, pp. 165-173. *http://dx.doi.org/10.3145/epi.2011.mar.06*

Rodríguez-Yunta, Luis (2012). "Control de autoridades, una herramienta desaprovechada en los sistemas de recuperación". *Anuario ThinkEPI*, v. 6, pp. 240-243.

IweTel, foro de información y debate de la biblioteconomía y la documentación, cumple 20 años

Fundada por Tomàs Baiget en 1993, *IweTel* es la lista pionera en español de los profesionales de las bibliotecas, documentación, bases de datos y sistemas de información en general.

Al principio se alojó en *Sarenet* y en 1998 pasó a *RedIRIS*. Posteriormente se han ido creado otras listas más especializadas como *Arxiforum* (archivos), *Bib-Med* (información bio-médica), *Bescolar* (bibliotecas escolares), *Incyt* (indicadores científicos), *Fidel* (recursos de internet), etc., pero *IweTel*, con más de 5.600 miembros, es la lista de referencia, el medio de comunicación básico y central para los profesionales de la información.

Funcionamiento

En la lista se cumple la conocida regla del 80/20 (el 80% de los mensajes los genera el 20% de los inscritos), o su reciente reformulación a 90, 9, 1%: el 90% de los inscritos son pasivos, casi nunca envían nada, el 9% (unos 360) participa alguna vez, y existe un 1% (50 personas) que genera la mayoría de mensajes. Con el aumento de inscritos y el número de mensajes (algunas semanas se distribuyen más de 100) fue necesario hacer la lista moderada, y en ello estamos los 4 firmantes, intentando aplicar nuestro sentido común para decidir cuáles se aprueban y cuáles no, y evitando los mensajes repetidos. Rechazamos alrededor de un 15-20%, lo cual a veces provoca quejas de sus autores. Para dirimir las dudas y para casos de conflicto, se creó un Consejo Asesor formado por veteranos de la lista, a quienes los moderadores pedimos consejo.

Objetivos

La lista cumple los dos objetivos básicos típicos: tablón de anuncios (conferencias, cursos, publicaciones, noticias) y foro de debates. Además se usa como sistema abierto de evaluación por pares (*open peer review*) de las notas que los miembros del think tank *ThinkEPI* envían periódicamente a la lista para su pública crítica y discusión. Esas notas y los principales mensajes que generan se publican cada año re-editados en el *Anuario ThinkEPI* de la editorial *EPI SCP*.

Las listas siguen valiendo

Con los cambios tecnológicos habidos a lo largo de estos años y, más recientemente, con las nuevas plataformas web 2.0, se ha planteado muchas veces si las listas de correo se han hecho "obsoletas". La verdad es que pensamos que una lista sigue siendo el medio ideal de comunicación de una comunidad profesional: rápida, limpia, discreta y eficaz, lejos de la faramalla de las redes sociales, también muy interesantes y útiles pero para otras cosas.

Más información e inscripciones:
http://www.rediris.es/list/info/iwetel.html

Javier Leiva-Aguilera (*Catorze.com*), **Paco López-Hernández** (*Universidad Carlos III de Madrid*), **Isabel Olea** (*Universidad de León*) y **Tomàs Baiget** (*El profesional de la información*).