J.1. Control de autoridades, una herramienta desaprovechada en los sistemas de recuperación

Luis Rodríguez-Yunta

28 noviembre 2011

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

Resumen: Se defiende que el catálogo de autoridades, además de una herramienta profesional, puede tener un rol importante en la recuperación de información. Se utiliza como ejemplo para subrayar cómo bibliotecas y bases de datos documentales no han sabido o no han logrado poner en juego todo su potencial en la sociedad del conocimiento.

Palabras clave: Recuperación de información, Autoridades, Lenguajes documentales, Desambiguación.

Title: Authority control, a wasted tool in recovery systems

Abstract: The author argues that the list of authorities, as well as a professional tool, may represent an important role in information retrieval. The purpose of this paper is to offer an example to highlight how libraries and document databases have failed or have not yet reached their full potential in the knowledge society.

Keywords: Information retrieval, Authorities, Documentary languages, Disambiguation.

Búsquedas más precisas

El catálogo de autoridades no parece desde luego un tema de actualidad. El propósito de este artículo no es mostrar nuevas aplicaciones de esta herramienta tradicional, sino por el contrario utilizarlo de ejemplo para subrayar cómo bibliotecas y bases de datos documentales no han sabido o aún no han logrado poner en juego todas sus potencialidades en la sociedad del conocimiento.

Aparentemente el fichero de autoridades es exclusivamente una herramienta técnica, en unos casos interno y en otros de demanda internacional, como los elaborados por la *Library of Congress* o las diferentes bibliotecas nacionales, pero siempre enfocado a la profesión. Su difusión en los sitios web de las bibliotecas a menudo se presenta como una opción de consulta independiente del catálogo. Cuando se aplica a la recuperación se limita a la búsqueda a través de los índices de materias, no en el formulario principal de consulta. Sin embargo, su utilidad podría ser mucho mayor para orientar al usuario y desambiguar términos dentro de la búsqueda libre. De igual modo, la filosofía del control de autoridades está presente en recursos terminológicos como los tesauros, y podría aplicarse en el mismo sentido para la búsqueda libre dentro de las bases de datos documentales.

La desambiguación es una cuestión clave en la recuperación de información. Numerosas palabras cambian de sentido según el contexto, y es imprescindible aclarar con el usuario qué sentido le otorga al buscar un término. El concepto es clave para el desarrollo de la web semántica, pero su aplicación es dificultosa, exigiría por ejemplo el uso sistemático de un sistema de metadatos de materia en los recursos web (**Granados-Colillas**, 2009). También es un concepto clave para los sistemas de inteligencia artificial, un campo con varias líneas de investigación para informáticos y lingüistas desde 1950 (**Sanderson**, 2000). El problema está aún lejos de resolverse en los programas de procesamiento de lenguaje natural, y una de las estrategias que se proponen radica en su aplicación selectiva, ya que sólo deben ser desambiguadas algunas palabras (**Perea-Ortega** *et al*., 2011).

A niveles sencillos y prácticos, la desambiguación se aplica con éxito en sistemas concretos desde los inicios de internet. Nadie entendería un buscador de compra de vuelos donde al solicitar que quiero ir de Miami a Barcelona el programa me ofreciera indistintamente viajes a España y otros a Venezuela en la misma respuesta. La solución es clara, cuando se teclea un término ambiguo, el sistema obliga a que el usuario se defina, mediante una tabla de valores válidos, pero que también se puede enriquecer con sinónimos y

relaciones. Y funciona perfectamente, porque hay un campo de búsqueda con una utilidad concreta que se contrasta con una tabla construida con coherencia.

Otro ejemplo de uso generalizado donde funciona perfectamente la desambiguación es *Wikipedia*. Su sistema de recuperación no intenta localizar simplemente registros que contengan la palabra tecleada, sino que devuelve al usuario una lista de valores posibles que responden a su búsqueda, basada exclusivamente en las entradas principales de la enciclopedia y sus relaciones de equivalencia (use-usado por). Y esto ocurre sólo si hay necesidad cuando hay varias opciones, no se obliga a este paso por defecto, pero tampoco se mezclan resultados de dudosa pertinencia en la respuesta al usuario.

Por el contrario, si busco Rubén Darío en el catálogo de la *Biblioteca Nacional de España*, en la respuesta se mezcla "Azul" con un título tan poco poético como "Estado de bienestar y trabajo social" (de un tal Rubén Darío Torres). ¿Sería mucho pedir que el opac me preguntase a qué Rubén Darío me refiero? Si no admitimos un fallo de este tipo en un buscador de vuelos, ¿por qué admitirlo en un catálogo bibliográfico, que además sí dispone de herramientas para ofrecer mayor precisión?

Efectivamente, es posible localizar las obras de y sobre el poeta Rubén Darío con una eficacia del 100%, pero exclusivamente mediante la consulta de los índices y la selección de la entrada adecuada. La herramienta existe en la mayor parte de los recursos documentales, salvo en las bases de sumarios, pero en ocasiones no se ofrece al usuario o bien queda oculta y no va a ser utilizada. Por ejemplo en las bases documentales del *CSIC* (*ISOC*, *ICYT*, *IME*) sí se ofrece la consulta de índices por campos, la única opción que puede resolver eficazmente las búsquedas que tienen problemas de ambigüedad. Pero se sabe que el usuario apenas utiliza este recurso e incluso algunos profesionales opinan que habría que quitar estos índices porque también muestran los errores tipográficos o las inconsistencias que intentamos evitar pero no siempre conseguimos corregir.

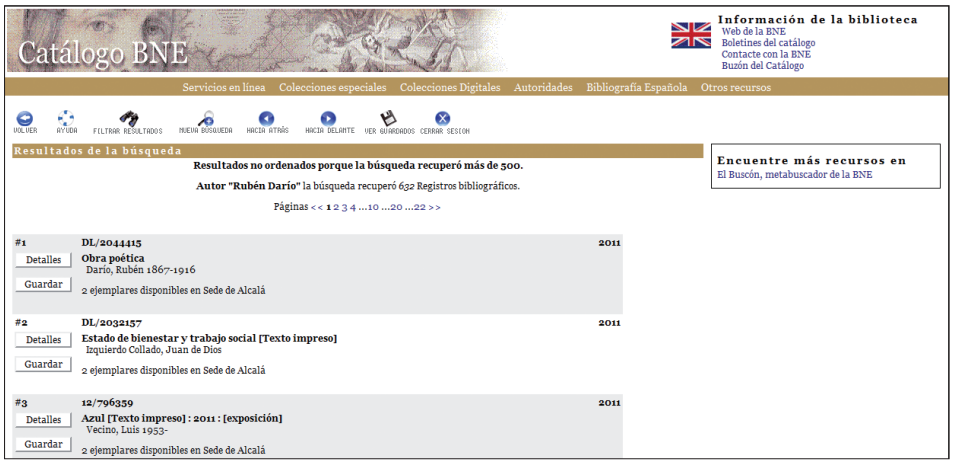
Poner en valor las listas de autoridades

Durante los últimos años es recurrente que la profesión convoque jornadas y conferencias para hablar de innovación y adaptación continua al cambio. Sin embargo, hay dos factores que dificultan que estos principios pasen de meras etiquetas a realidades tangibles:

"El sistema de recuperación de los opacs y bdds documentales no debe compararse con *Google* sino con el buscador de vuelos o con *Wikipedia*"

1. Las bibliotecas y centros de documentación en general no son creadores sino consumidores de software. Se limitan a instalar un paquete con un diseño meramente "razonable" y conviven con ese nivel para siempre. Carecen de capacidad para la mejora continua que muestran otros sectores. No pueden perseguir la excelencia porque no viene en el diseño estándar del programa. Y el resultado es una multitud de presentaciones casi idénticas que repiten las mismas limitaciones, sin que los errores de uno puedan servir de alerta para otros.

2. Se confunde el modelo con el que debe compararse el sistema de recuperación de los catálogos y bases de datos documentales. Se ha buscado asemejarse a los buscadores genéricos como *Google*, cuando debería compararse con sistemas de objetivo específico como el buscador de vuelos o *Wikipedia*.



En general, el usuario de un recurso bibliográfico no busca "cualquier cosa que contenga X". Utiliza un catálogo o una base de datos para un número de posibles utilidades muy reducido que pueden ofrecerse en un menú sencillo. ¿Por qué entonces la insistencia en ofrecer un formulario simple que parece servir para todo y en consecuencia ignora los objetivos del usua-

rio? Por supuesto, el impacto de *Google* sobre los hábitos de búsqueda es incuestionable. Pero ello no quiere decir que sólo puedan ser apreciados aquellos programas que muestren un funcionamiento similar.

¿Para qué parecerse a un producto si el usuario no va a preferir otro por el criterio de similitud? ¿No sería preferible no parecerse y reforzar utilidades diferenciadoras? ¿Es que algún usuario utiliza *Google* para contratar un viaje, alquilar un piso o buscar pareja? Lo utiliza para localizar páginas diseñadas para estos usos específicos pero una vez dentro de ellas no espera que su diseño sea el mismo que el del buscador, sólo desea un producto que le de precisión y eficacia en la selección de resultados.

En el contexto actual, la pertinencia en los primeros registros que se muestran es fundamental para ofrecer una imagen de eficacia. Los buscadores de objetivo genérico han procurado solucionar este problema en base a algoritmos complejos de ordenación que buscan atinar con las demandas más generalizadas de un término (que pueden variar según el país).

Los productos bibliográficos, o más bien, los programadores de software para vender a las bibliotecas y centros de documentación, han añadido este concepto de ordenación por relevancia, como una opción que podemos implementar por defecto o no, pero que tiene bastante poco que ver con aprovechar las herramientas de normalización e indización con las que cuentan estas instituciones. Y además sus resultados son bastante menos eficaces que en *Google*, porque se carece de la capacidad de adaptación y mejora continua del programa. La búsqueda por índices y el control de autoridades, que podrían garantizar la máxima eficacia, han quedado marginados en el diseño.

Las listas de autoridades se citan entre las herramientas susceptibles de incorporarse en proyectos de *open data* y *linked open data* (**Peset**; **Ferrer-Sapena**; **Subirats-Coll**, 2011). Podría darse la paradoja de que haya aplicaciones ajenas al sector de la biblioteconomía y documentación que implementen estas fuentes bajo la filosofía de compartir datos, y sin embargo los recursos bibliotecarios y documentales continúen desaprovechando su potencial.

Referencias bibliográficas

Granados-Colillas, Mariàngels. "Tratamiento de los metadatos de contenido en la web semántica o cuando la lógica se hace evidente". En: *Nuevas perspectivas para la difusión y organización del conocimiento. IX Congreso ISKO-España*. Valencia: Universidad Politécnica, 2009, pp. 1043-1054.
http://eprints.rclis.org/handle/10760/13223

Perea-Ortega, José M.; **Martínez-Santiago, Fernando**; **García-Cumbreras, Miguel A.**; **Montejo-Ráez, Arturo**. "Desambiguación de términos basada en IDF aplicada a recuperación de información". *Procesamiento del lenguaje natural*, 2011, marzo, n. 46, pp. 99-106.
http://sinai.ujaen.es/sepln/ojs/ojs-2.3.5/index.php/pln/article/view/849/704

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

Sanderson, Mark. "Retrieving with good sense". *Information retrieval*, 2000, n. 2, pp. 47-67.
http://dx.doi.org/10.1023/A:1009933700147

* * *

Nuevo papel de las bibliotecas ante el *linked open data*

Juan-Antonio Pastor-Sánchez

La nota de **Luis Rodríguez-Yunta** me sugiere un nuevo papel de las bibliotecas, abriendo sus catálogos (incluyendo los de autoridades) de manera que se incorporen a la realidad de la web semántica en general y de *linked open data* en particular. También me ha hecho reflexionar sobre la carencia de software para la gestión de autoridades y vocabularios, totalmente libre, fácil de instalar, configurar e integrar de manera sencilla y transparente con el resto de aplicaciones en la biblioteca, al tiempo que se usen en LOD dentro del campo de las bibliotecas e incluso de otras unidades de información.

Es interesante el documento sobre *Datos enlazados bibliotecarios: conjuntos de datos, vocabularios controlados y conjuntos de elementos de metadatos*, elaborado por el *W3C Incubator*. **Tomás Saorín** y yo hemos hecho una traducción[1]. Se trata de una versión provisional, que junto con *Informe final de datos enlazados bibliotecarios*[2] y con casos de uso[3] suponen documentos esenciales para entender un nuevo contexto que remarca la nota de **Rodríguez-Yunta**.

La web semántica ofrece muchas posibilidades, pero también existen soluciones adecuadas para la publicación de catálogos de autoridades. Skos puede usarse para ello (complementándolo cuan-

do sea necesario mediante owl) sin necesidad de diseñar nuevas ontologías, y aprovechando sus posibilidades de mapeado.

1. http://skos.um.es/Incubator/lld/XGR-lld-vocab dataset

2. http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025

3. http://www.w3.org/2005/Incubator/lld/XGR-lld-usecase-20111025

pastor@um.es

Proyectos y desarrollo de aplicaciones

Xavier Agenjo

Todo lo que señala **Juan-Antonio Pastor-Sánchez** es cierto, si bien se pueden introducir algunos matices, posibles por la existencia del *Europeana data model* al que suele denominarse el "sabor" europeo de *linked open data*. Hay que prestar cierta atención a ese sabor porque *Europeana* es la *Acción clave n° 15* de la *Agenda digital europea* y tiene unas repercusiones enormes sobre todo el trabajo que se haga en archivos, bibliotecas y museos en España. Considerando que no se atraviesan momentos de gran abundancia económica y sí de fuertes reducciones presupuestarias, convendría tener muy presente esta normativa europea.

En el documento que se cita de *linked open data* aparece, entre los 12 casos de estudio, un proyecto de la *Fundación Ignacio Larramendi* denominado *Biblioteca virtual de polígrafos*, traducido como *Polymath Virtual Library*. Esta biblioteca es capaz de intercambiar datos siguiendo el *Europeana data model* y se ajusta desde luego a *linked open data*, aunque éste es un camino que hay que recorrer, enriqueciendo y agregando datos –por no decir de un modo más preciso "objetos digitales"– que conforman las bases de datos. Con este mismo programa, en diversas versiones, hay en España más de 30 implementaciones que contribuyen con aproximadamente un 40% al *Europeana linked open data pilot project*. Seguramente no es por casualidad.

En los últimos tiempos hemos avanzado mucho en la implementación y desarrollo de esa normativa, y muy en concreto en un proyecto que hemos denominado *Biblioteca virtual de la Escuela de Salamanca*[1]. Como podréis ver por el ejemplo que adjunto con su enlace, el registro de autoridad puede transformarse desde el formato ficha donde se puede identificar inmediatamente los enlaces a VIAF o a la *DBpedia*, una estructura MARC 21 etiquetada que permite apreciar los campos RDA que MARC ha ido definiendo; por supuesto, el registro se transforma automáticamente en DC RDF y de forma transparente alimenta el correspondiente repositorio OAI-PMH.

También permite la exportación en MARC XML y en el ya citado *linked open data* con el sabor ese europeo. De paso diré también que permite exportar tanto en el formato de autoridades de archivos *Encoded archival context* como en *Isaar* (*CPF*).

A mi modo de ver, no es especialmente complejo el desarrollo de aplicaciones que utilicen como modelo de datos las especificaciones funcionales y estándares que prescriben *Europeana* o el *W3C Library Linked Data Incubator Group*. Sencillamente, hay que ponerse a ello.

1. http://www.larramendi.es/i18n/consulta_aut/registro.cmd?control=POLI20090012677

agenjo@ono.com