Registros de autoridades, enriquecimiento semántico y *Wikidata*

Authority records, semantic enrichment, and *Wikidata*

Xavier Agenjo-Bullón y Francisca Hernández-Carrascal

Agenjo-Bullón, Xavier; **Hernández-Carrascal, Francisca** (2018). "Registros de autoridades, enriquecimiento semántico y *Wikidata*". *Anuario ThinkEPI*, v. 12, pp. 361-372.

https://doi.org/10.3145/thinkepi.2018.61

Publicado en *IweTel* el 11 de diciembre de 2017

Resumen: La nota da cuenta de los procesos que realizan archivos, bibliotecas y museos para enriquecer semánticamente los registros de autoridad con distintos fines, especialmente para mejorar su usabilidad y para fortalecer la búsqueda y recuperación de información con nuevas funcionalidades. Se describen las tareas que habitualmente se engloban dentro del término enriquecimiento semántico: la normalización de datos, la reconciliación y vinculación con recursos *linked open data* externos, la extracción de datos adicionales y la publicación de datos. Se plantea que enriquecer y vincular grandes cantidades de datos no puede realizarse sino por procedimientos automáticos, como se ha hecho en la *Biblioteca Virtual de Polígrafos* para el enriquecimiento semántico de los registros de autoridad de polígrafos. Por último se analiza el papel que ha adquirido *Wikidata* como uno de los recursos *linked open data* más utilizados a nivel mundial, y que a su vez se ha convertido en un gran fichero de autoridades para todos los proyectos *Wikimedia*.

Palabras clave: Control de autoridades; Enriquecimiento semántico; Reconciliación de datos; Registros de autoridad; Datos abiertos vinculados; LOD; Datos abiertos enlazados.

Abstract: This article describes the processes carried out by archives, libraries, and museums to semantically enrich authority records with the aim of improving their usability while building new functionalities for search and retrieval. The tasks that are usually included within the term semantic enrichment are: data cleaning, data reconciliation to external linked open data resources, extraction of additional data, and data publication. However, reconciliation and enrichment of large volumes of data can only be done by automatic means; for example, the semantic enrichment of polymath authority records by the Spanish *Polymath Virtual Library*. Finally, we analyze the role that *Wikidata* has as one of the most widely used linked open data resources and authority file for all *Wikimedia* projects.

Keywords: Authority control; Semantic enrichment; Data reconciliation; Authority records; Linked open data (LOD).

1. Introducción

Aunque hay alguna bibliografía sobre los registros de autoridad como fuente de información, hay bastante menos sobre el enriquecimiento semántico con vocabularios de valores, a pesar de ser una práctica recomendada en el *Informe Final del Grupo Incubador de Datos Vinculados de Bibliotecas* (**Baker** *et al.*, 2011) publicado el 25 de octubre de 2011. Pero, además, recientemente se ha cumplido el quinto aniversario de *Wikidata*, (**Lih**; **Fernández**, 2017) que ha llegado a ser un recurso de información semántica fundamental. Es más, se podría decir que se está convirtiendo en el fichero de autoridades por antonomasia. Esta nota quiere dar cuenta de ambas cuestiones que ilustraremos con la práctica de la *Biblioteca Virtual de Polígrafos* de la *Fundación Ignacio Larramendi*, implementada sobre el software *Digibib*, desarrollado por la empresa filial de la *Fundación*.

Indudablemente, desde el punto de vista de la satisfacción del usuario la posibilidad de enriquecer semánticamente los metadatos mejora su usabilidad, su consulta y abre numerosas vías de navegación. Por ello, la vinculación con los vocabularios de valores disponibles en un dominio concreto no debe ser una posibilidad más o menos realizable, sino una técnica que debe incorporarse a las tareas de control de autoridades.

Ese enriquecimiento semántico confiere una gran calidad a los registros, pero se impone una visión biblioteconómica del proceso. ¿Es factible vincular todos los registros de un catálogo con diferentes vocabularios de valores? Es evidente que para actualizar un registro de autoridad hay que consumir un tiempo T, para vincular el registro a cada vocabulario de valores hay que consumir un tiempo T1, T2, T3... Tn, puesto que son muchos los vocabularios de valores disponibles y, gracias a Dios, cada vez hay más.

En la *Fundación Ignacio Larramendi* y en *Digibís* hemos puesto en práctica un procedimiento para que todos los registros de autoridad estén vinculados por procedimientos automáticos, aunque siempre hay que validar casos dudosos y definir estrategias para contrastar los datos procedentes de distintas fuentes. Fruto de esas actividades ha emergido la autoridad de *Wikidata*.

2. Las tareas de enriquecimiento semántico en archivos, bibliotecas y museos

Entre los casos de uso sobre los que se elaboró el *Informe Final* del *W3C-LLD*, el correspondiente al 'enriquecimiento de datos de autoridades' se describía así:

> "Un bibliotecario quiere enriquecer un registro de autoridad de su institución (por ejemplo, una persona) con información adicional. Como primer paso, el bibliotecario debe buscar entidades de autoridad equivalentes en el conjunto de datos de uno o más proveedores de datos. A continuación, el bibliotecario debe identificar la entidad equivalente en el resultado de la búsqueda. En caso de una identificación exitosa, ambas representaciones deben estar alineadas y el bibliotecario debe decidir qué características y/o relaciones son importantes para el enriquecimiento del registro local"[1].

Uno de los ejemplos más importantes, por el volumen, variedad de los datos y de los vocabularios, de cómo se están asumiendo y desarrollando las nuevas tareas de enriquecimiento semántico lo proporciona *Europeana*, que consideró inicialmente el enriquecimiento de datos como un experimento, pero finalmente lo incorporó a su estrategia (**Stiller**; **Isaac**; **Petras**, 2014). En 2015 la *Task force on enrichment and evaluation* de *Europeana* (**Isaac**, *et al.*, 2015) publicó el resultado de su trabajo en un informe del que extraemos los siguientes puntos:

- para las instituciones culturales, el enriquecimiento de metadatos se ha convertido en una forma de superar los problemas de calidad de los datos, al mismo tiempo que proporciona más información contextual y multilingüe (p. 3);
- generalmente, la tarea de enriquecimiento de metadatos se puede describir como un proceso que mejora los datos sobre un objeto, agregando nuevas declaraciones sobre el objeto descrito. El término "enriquecimiento" puede usarse para referirse al proceso (por ejemplo, la aplicación de una herramienta de enriquecimiento o su resultado), o a los nuevos metadatos generados al final del proceso. La estrategia de enriquecimiento se refiere a todos los componentes de los flujos de trabajo y a los procesos que determinaron estos componentes (p 5).

Algo muy similar relata **Morillo-Calero** (2014) sobre las actividades de enriquecimiento semántico realizadas por la *Digital Public Library of America*[2].

A partir de las distintas experiencias que se han ido poniendo en práctica, y de la nuestra en particular, podemos decir que la expresión "enriquecimiento semántico" engloba distintas tareas que en la mayoría de las ocasiones se corresponden con procesos reiterativos[3] difícilmente distinguibles por sí mismos:

- normalización de los datos (por no decir limpieza de datos), lo que incluye correcciones

ortográficas, de puntuación, de uso de mayúsculas y minúsculas, detección de duplicaciones, y en muchos casos la homogeneización de los valores de diferentes campos, especialmente de aquellos con los que se realizará la búsqueda y correspondencia en vocabularios de valores externos;

- vinculación con fuentes *linked open data*, proceso al que se denomina específicamente "reconciliación de datos", lo que es una traducción literal de la expresión inglesa *reconciliation*, y que engloba la tarea misma, el proceso informático de buscar una entidad equivalente en esas fuentes de forma automática, puntuar el grado de correspondencia entre los valores obtenidos y extraer el URI del recurso coincidente;
- extracción de datos adicionales o complementarios, especialmente los multilingües, de las fuentes con las que se han vinculado los datos y que llevan a la creación de nuevas entidades, propiedades o atributos;
- publicación de los datos bibliotecarios vinculados y enriquecidos en la Web, en general, y en sistemas de información no bibliotecarios de amplio uso (por ejemplo, y como veremos más adelante en *Wikidata*);
- mantenimiento de los datos actualizados ya sea por la incorporación de nuevos registros o por la publicación de nuevas fuentes *linked open data*.

En general, la transformación y publicación de datos en *linked open data* implica su enriquecimiento semántico y constituye uno de los retos a los que se enfrentan los archivos, bibliotecas y museos en la actualidad. Como podemos ver por el desglose de tareas anterior, el problema no es ya tanto el conocimiento o la utilización de *linked open data* sino el diseño de los flujos de trabajo que permitan la realización del control de autoridades en un entorno RDF y *linked open data*[4].

En los resultados de las encuestas de 2014 y 2015 lanzadas por *OCLC*[5] para evaluar los proyectos y servicios *linked open data* publicados por **Smith-Yoshimura** (2016), se mencionan dos grupos de comentarios que nos interesan especialmente:

> "el reposicionamiento del conocimiento bibliotecario para proporcionar acceso a los recursos en la web semántica como una red de actividades";
>
> "la transformación de las organizaciones para cambiar la gestión de los catálogos MARC a RDF y *linked open data* por medio de nuevos flujos de trabajo, en un mundo plagado de datos duplicados"[6].

Para que las instituciones culturales puedan consumir y producir datos abiertos vinculados necesitan enriquecer y vincular grandes cantidades de datos, lo que difícilmente se podrá realizar por procedimientos manuales que requieren una enorme inversión en tiempo y recursos humanos. Para ello, aparte de los desarrollos específicos que puedan realizarse, existen diferentes programas (**Patnab**, 2015) que permiten realizar parcial o totalmente todas las tareas involucradas en el proceso de enriquecimiento, de forma masiva y semiautomática. Es obligado aclarar que masivo y semiautomático no quiere decir que sean procesos que puedan realizarse de forma desatendida sin la mediación de profesionales especializados, sino que tienen una productividad que simplemente hace viable lo que manualmente es imposible.

"La transformación y publicación de datos en *linked open data* implica su enriquecimiento semántico y constituye uno de los retos a los que se enfrentan los archivos, bibliotecas y museos en la actualidad"

3. Enriquecimiento semántico en la *Biblioteca Virtual de Polígrafos*

Como se anunció en el artículo *Data aggregation and dissemination of authority records through linked open data in a European context* (**Agenjo-Bullón**; **Hernández-Carrascal**; **Viedma**, 2012), presentado previamente en el Congreso de *IFLA* de 2011 en Puerto Rico (**Agenjo-Bullón**; **Hernández-Carrascal**; **Viedma**, 2011), la *Biblioteca Virtual de Polígrafos* cuenta con un muy bien definido fichero de autoridades en formato MARC21, que de forma transparente y dinámica puede transformarse según el esquema XML de *Europeana Data Model* (EDM) y alimentar su repositorio OAI-PMH; o bien devolver la descripción correspondiente en EDM RDF a través de la negociación de contenido (**Thereaux**, 2006).

Los registros en formato MARC21 están vinculados, por medio del campo 024, con una pluralidad de vocabularios de valores, entre los que se cuentan *VIAF*, *datos.bne.es*, *DBpedia*, *Wikidata*, y otros. Ya en 2014 la *Biblioteca Virtual de Polígrafos* figura como caso de estudio de EDM (**Charles**, 2014), así como fue caso de estudio del *W3C LLD*.

Todo ello puede verse en el registro de autoridad de Miguel de Unamuno.
https://goo.gl/rXetwf

En efecto, entre las fuentes utilizadas para vin-

cular a los registros de autoridad de polígrafos se han 'reconciliado', además de la propia *DBpedia*, las *dbpedias* de los ámbitos culturales de los polígrafos hispanoamericanos, portugueses y brasileños, y de todos los idiomas oficiales de España, es decir la *DBpedia* en español, la *DBpedia do galego* y el *DBpediako euskarazko kapitulua*. Por desgracia no hemos sido capaces de localizar la *DBpedia* en catalán.

http://wiki.dbpedia.org
http://wiki.dbpedia.org/about/language-chapters
http://es.dbpedia.org
http://gl.dbpedia.org/wiki
http://eu.dbpedia.org

Esta técnica de vincular, y enriquecer, los registros de autoridad con la *DBpedia* es muy útil puesto que nos proporciona automáticamente, además de otros posibles datos, una pequeña biografía, que se consigna en el campo 678 del formato MARC21. Es evidente que para Unamuno esa información se queda muy corta, pero para miles y miles de autores constituye una fuente de información importantísima… cuando no se dispone de ninguna otra.

Hay que señalar que, en el proceso de enriquecimiento de los registros de autoridad de polígrafos, *Wikidata* se ha revelado como una fuente de datos fundamental por diferentes motivos, pero especialmente por la gran cantidad de datos estructurados, en constante mejora, por su estrecha relación con *Wikipedia* y otros proyectos *Wikimedia* y porque a su vez está relacionada con muchos otros vocabularios de valores, externos a *Wikipedia*, pero propios de archivos, bibliotecas y museos.

Sobre el ejemplo de la descripción de Unamuno en la *Biblioteca Virtual de Polígrafos* y los datos disponibles en

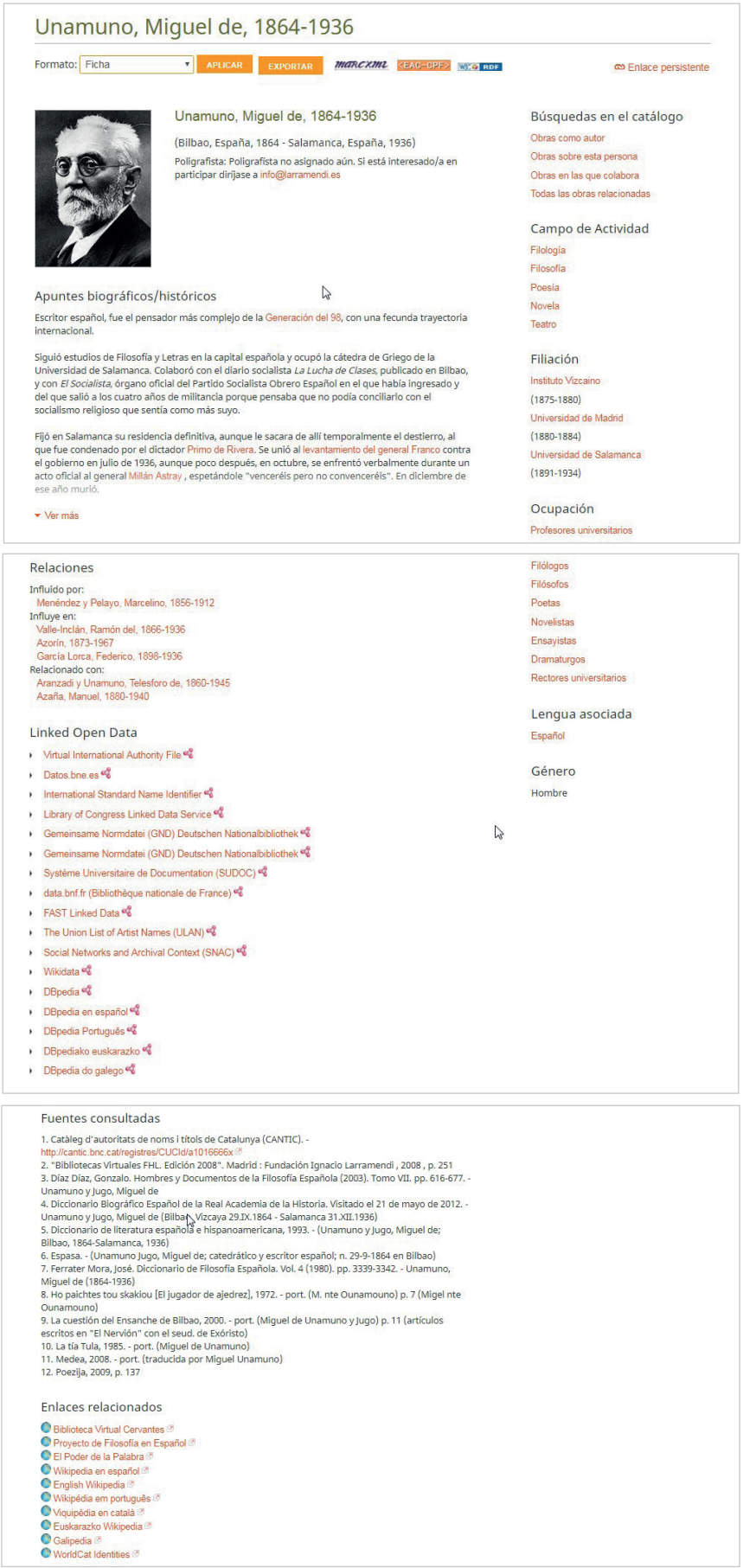

Figura 1. Miguel de Unamuno en la *Biblioteca Virtual de Polígrafos* *https://goo.gl/rXetwf*

VIAF, *DBpedia*, *Wikidata* y la propia *Biblioteca Virtual de Polígrafos* mostramos un resumen comparativo, y no exhaustivo, de los datos que más nos interesan de cada una de las fuentes (tabla 1). En esta misma tabla se han coloreado las filas para indicar los subprocesos del enriquecimiento semántico en el que cada tipo de dato participa principalmente. Las filas en blanco se refieren a propiedades que no se han utilizado.

- Reconciliación de datos (filas de color rosa claro): el nombre y los nombres alternativos, se utilizan para poder establecer equivalencias de nombres entre unas fuentes u otras, que es a fin de cuentas el significado del término reconciliación. Para ello muchas fuentes *linked open data* ofrecen también distintos servicios de reconciliación, cada uno con sus peculiaridades de funcionamiento (a través de APIs o de *Sparql endpoints*) y estructuración de los datos. Incluso la reconciliación de datos puede realizarse por medio de la descarga directa de ficheros en formatos estructurados, desde CSV a RDF, lo que amplia notablemente el número de recursos reutilizables dado que muchos conjuntos de datos todavía solo se publican en formatos CSV o *Excel*.
- Vinculación con fuentes *linked open data* (filas de color azul claro): una vez que se ha establecido la correspondencia de una descripción con otra, y se está seguro de que no se trata de un falso positivo (en cuya elucidación pueden y deben intervenir otras propiedades además del nombre), es posible obtener los URIs que identifican esas descripciones e incorporarlos a nuestros datos.
- Datos adicionales que pueden extraerse (coloreados en verde): son todas aquellas aseveraciones que nos puede interesar extraer. Obviamente queda al criterio y conocimiento del catalogador determinar el valor que puede atribuirse a cada fuente, o a cada dato en particular, y la conveniencia de utilizarlo o no. Es necesario conocer a fondo la estructura de datos de cada fuente y el uso que se hace de esa estructura en las descripciones, ya que pueden coexistir variantes y darse la circunstancia de que instancias de una misma clase puedan tener propiedades distintas para el mismo objetivo descriptivo. Por ejemplo, algunas personas pueden tener como lugar de nacimiento el país únicamente, mientras otras ofrecen la localidad concreta. Al igual que nuestros ficheros de autoridades las fuentes externas *linked open data* también están en constante mejora.

3. *Wikidata*: el control de autoridades en fuentes no bibliotecarias

Desde la publicación del *Informe Final* en 2011 han tenido lugar en la Web una serie de hitos que son reflejo de los diferentes aspectos funcionales y técnicos de los datos abiertos vinculados:

1) La transformación masiva de datos y metadatos en datos abiertos vinculados. La última "nube *linked open data*" publicada en agosto de 2017 (**Abele** *et al.*, 2017) muestra que en 10 años el número de conjuntos de datos publicado se ha incrementado en un 4.153,57%.

2) La extensión del proceso de vinculación de datos a un número cada vez más amplio de recursos *linked open data*. Según la *Linked Data Survey* de *OCLC* de 2015, los conjuntos más utilizados para la vinculación y enriquecimiento de recursos son:

- *Virtual International Authority File* (*VIAF*). *https://viaf.org*
- *DBpedia. http://wiki.dbpedia.org*
- *GeoNames. http://www.geonames.org*
- *Library of Congress Linked Data Services. http://id.loc.gov*
- *Getty's Art and Architecture Thesaurus. http://www.getty.edu/research/tools/vocabularies/aat*

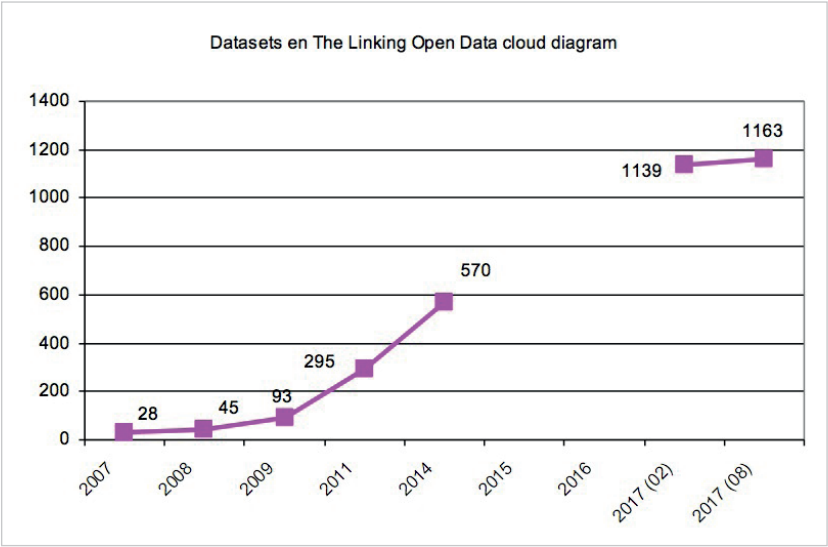

Figura 2. Evolución del número de conjuntos de datos en la nube *linked open data*

Tabla 1. Descripción de Unamuno en la *Biblioteca Virtual de Polígrafos*. Datos disponibles en *VIAF*, *DBpedia*, *Wikidata* y la propia *Biblioteca Virtual de Polígrafos*

Datos	*VIAF*	*DBpedia*	*Wikidata*	*Biblioteca Virtual de Polígrafos*
URI	*http://viaf.org/viaf/88844103*	*http://dbpedia.org/resource/Miguel_de_Unamuno*	*http://www.wikidata.org/entity/Q185085*	*http://www.larramendi.es/aut/POLI20090015128*
Nombre	No existe un único nombre, agrupa todos los nombres de los registros que forma el clúster		Miguel de Unamuno	Unamuno, Miguel de, 1864-1936
Formas alternativas del nombre	142		56	25
Instancia de	Nombre de persona	65 categorías (entre las que se incluyen *Thing*, *Person*, *Agent*… *Philosopher*, *People from Bilbao*…)	Ser humano	
Relaciones	6 (5XX)	*Influenced* (3), *Influenced by* (5)		Influido por Influye en Relacionado con (5XX)
Coautores	10			
Género			Hombre	Hombre
País de ciudadanía / país relacionado			España	España
Fecha de nacimiento		1864-09-29	29 septiembre de 1864	1864
Lugar de nacimiento		Spain, Bilbao, Biscay	Bilbao	Bilbao
Fecha de defunción		1936-12-31	31 diciembre 1936	1936
Lugar de defunción		Salamanca, Spain, Province of Salamanca	Salamanca	Salamanca
Idiomas utilizados			Español y vasco	Español
Escuela filosófica		Scholasticism, Existentialism, Positivism, Platonism		
Campo de actividad				Filología, Filosofía, Poesía, Novela, Teatro
Ocupación			Poeta, filósofo, escritor, ensayista, novelista, crítico literario, profesor universitario, dramaturgo	Profesores universitarios, Filólogos, Filósofos, Poetas, Novelistas, Ensayistas, Dramaturgos, Rectores universitarios
Miembro de / Instituciones relacionadas			*Real Academia Española*, *Sociedad de Amigos de Portugal*	*Instituto Vizcaino* (1875-1880), *Universidad de Madrid* (1880-1884), *Universidad de Salamanca* (1891-1934)
Formación		Alma mater: *Universidad Complutense de Madrid*	*Universidad Central*	
Biografía		Abstract	Enlaces a las *wikipedias*	Biografía de elaboración propia y resumen extraído de Wikipedia (a través de DBpedia)
Identificadores externos vinculados	*VIAF*, ISNI, LC, GND, *BNF*, *BNE*, *CANTIC* y 30 identificadores más.	20, (*VIAF*, GND, *Wikidata*, y otros; 13 enlaces a *DBpedias* en distintos idiomas)	*VIAF*, ISNI, LC, GND, *BNF*, *BNE*, *CANTIC*… y 44 identificadores más	VIAF, ISNI, LC, GND, BNF, BNE, Wikidata, DBpedia y 9 identificadores más. Enlaces
Enlaces externos	52 enlaces a *Wikipedias* en distintos idiomas		52 enlaces a *Wikipedias* en distintos idiomas, 23 a *Wikiquote*, 3 a *Wikisource*	*Wikipedia* en español English *Wikipedia* *Wikipédia* em português *Viquipèdia* en català *Euskarazko Wikipedia* *Galipedia*… 4 enlaces más, entre ellos *WorldCat Identities*

- *FAST* (*Faceted Application of Subject Terminology*) *Linked Data* (*OCLC*).
  http://experimental.worldcat.org/fast
- *WorldCat.org* (*OCLC*)
  http://www.worldcat.org
- *data.bnf.fr* (*Bibliothèque nationale de France*).
  http://data.bnf.fr
- *Deutsche National Bibliothek Linked Data Services.*

3) Las nuevas tareas y flujos de trabajo en archivos, bibliotecas y museos derivadas del enriquecimiento semántico, que ya hemos mencionado.

"Para que las instituciones culturales puedan consumir y producir datos abiertos vinculados necesitan enriquecer y vincular grandes cantidades de datos, lo que difícilmente se podrá realizar por procedimientos manuales que requieren una enorme inversión en tiempo y recursos humanos"

4) La extensión del control de autoridades a ámbitos no bibliotecarios.

Entre estos hitos merece mención especial uno de los hechos más importantes acontecidos en la biblioteconomía mundial en los últimos años, como es la extensión del control de autoridades al ámbito de la *Wikipedia*. Este proceso ha supuesto la vinculación de grandes cantidades de datos entre los ficheros de autoridades de las instituciones culturales y la *Wikipedia*, y la generación de Wikidata (**Fredo** *et al.*, 2014) como la gran base de conocimientos multilingüe de los proyectos *Wikimedia*, que actúa también a modo de fichero de autoridades de *Wikimedia*.

En 2012 se lanzó la *Wikipedia:Authority control integration proposal*[7] para la utilización de los identificadores del *Virtual International Authority File* en la *Wikipedia*, especialmente en las biografías. Supuso también el desarrollo de un programa (*VIAFbot*) para la detección de los identificadores *VIAF* apropiados y su adicción a los artículos de *Wikipedia*, así como el refinamiento de la *Plantilla:Control de autoridades*[8] y su uso extensivo en cientos de miles de artículos (**Klein**; **Kyrios**, 2013). La *Plantilla:Control de Autoridades*, que se creó en 2009 en el marco de la *Deutschsprachige Wikipedia*, tiene como finalidad agregar automáticamente en la parte inferior de los artículos una serie de identificadores en diversas bases de datos, tomados de *Wikidata*.

"En el proceso de enriquecimiento de los registros de autoridad de polígrafos, *Wikidata* se ha revelado como una fuente de datos fundamental"

Siguiendo con nuestro ejemplo de Unamuno, al pie de la entrada de *Wikipedia* puede verse el citado marco para el control de autoridades (figura 3).

El interés de los ficheros de autoridad de las instituciones culturales para *Wikidata* queda claramente expresado en la entrada *Wikidata:WikiProject Authority control*[9], creada en 2015, que comienza con las siguientes líneas:

> "Wikidata pays a lot of tribute to authority control, linking to all kinds of datasets and databases with various IDs. The holy grail of every GLAM worker Sum of All People, with links to their Works is coming about!"

En la actualidad los artículos y páginas de *Wikipedia* que están relacionados con identificadores de ficheros de autoridad y otras fuentes han crecido notablemente y continúan haciéndolo. Si la propuesta inicial de interrelación entre *Wikipedia* y *VIAF* estaba centrada en las personas y en la *English Wikipedia*, este procedimiento se ha ido extendiendo gradualmente. La *Wikipedia* en español empezó a ofrecer esta información

Figura 3. Marco para control de autoridades en el pie de la entrada de Unamuno en *Wikipedia*.
https://es.wikipedia.org/wiki/Miguel_de_Unamuno

en 2016, y en la actualidad está presente en personas, entidades, obras y bienes culturales, entre otros. En *Wikipedia:Artículos con control de autoridades*[10] se agrupan los artículos en diferentes categorías, entre ellas:

- Artículos con control de autoridades de autores;
- Artículos con control de autoridades de enciclopedias o diccionarios;
- Artículos con control de autoridades de obras;
- Artículos con control de autoridades de bienes culturales, etc.

En la tabla anexa a esta nota se puede ver el resultado de una búsqueda en el servicio *Sparql* de *Wikidata* con la relación de las fuentes españolas utilizadas para el control de autoridades en *Wikidata*.

Las cifras de los conjuntos de datos que giran en torno a *Wikipedia* son impresionantes:

- 5.485.590 artículos en inglés y 1,356,482 en español
  https://stats.wikimedia.org/EN/Sitemap.htm
- 4,58 millones de descripciones en *DBpedia*
  http://wiki.dbpedia.org/about/facts-figures
- 38,6 en Wikidata.
  https://stats.wikimedia.org/wikispecial/EN/TablesWikipediaWIKIDATA.htm

Y no hay que olvidar que *Wikipedia* es el quinto sitio web del mundo y, por tanto, un instrumento de primera magnitud para los archivos, bibliotecas y museos.
https://www.alexa.com/topsites

En definitiva, la extensión del control de autoridades a la *Wikipedia*, *Wikidata* y *DBpedia*, y la vinculación de los datos de este recurso con los identificadores de ficheros de autoridad de archivos, bibliotecas y museos amplía las posibilidades de enriquecimiento de datos y hace posible utilizar estas fuentes para la extracción de datos no disponibles habitualmente en los ficheros de autoridades.

"En 2012 se lanzó la *Wikipedia:Authority control integration proposal* para la utilización de los identificadores del *Virtual International Authority File* en la *Wikipedia*, especialmente en las biografías"

Desde la creación de *Wikidata* en 2012, su quinto aniversario se produjo el 29 de octubre pasado[11], tanto los datos como las herramientas de las que dispone para crear descripciones totalmente multilingües, han hecho que su valor como fuente de datos abiertos vinculados esté al mismo nivel que *DBpedia*, o incluso superándola por lo que se aprecia en su creciente nivel de uso para el enriquecimiento semántico.

4. Conclusión

Como conclusión, las técnicas de enriquecimiento semántico se benefician de los procesos de reconciliación con diferentes vocabularios de valores, entre los que destacan ya no sólo las fuentes bibliotecarias, sino también fuentes no bibliotecarias como *Wikidata*. En su quinto aniversario está muy claro que se está convirtiendo en un recurso autorizado a nivel mundial, dentro de la web semántica.

Los bibliotecarios, archiveros y museólogos contamos además con la ontología de *Europeana Data Model*, lo que permite publicar fácilmente todos los datos citados, y si además se dispone de una aplicación informática apropiada para ello todos esos mapeos de información pueden ser dinámicos, sin necesidad de duplicar los datos.

Los resultados de las tareas que se han reflejado en esta nota están disponibles en la *Biblioteca Virtual de Polígrafos* desde hace 8 meses, con la presentación de la *Biblioteca Virtual de Viajeros Científicos Ilustrados*, en cuya nota a esta edición digital[12] se daba cuenta de las novedades introducidas.
http://www.larramendi.es/vcilustrados/es/micrositios/inicio.do

Otros proyectos que ya han sido realizados y que aún no se han publicado incluyen el enriquecimiento semántico de instituciones y localidades con la geolocalización de las mismas por diferentes sistemas de coordenadas geográficas.

Sin embargo, aunque hemos hecho un gran esfuerzo en mejorar la visualización de los registros de autoridad es mucho lo que se puede hacer todavía, para lo cual es de gran ayuda el trabajo que están realizando otras instituciones en el mismo sentido.

Notas

1. *W3C.* "Use case authority data enrichment"
https://www.w3.org/2005/Incubator/lld/wiki/Use_Case_Authority_Data_Enrichment

2. **Morillo-Calero, María-Jesús** (2014). *Informe ALA Midwinter Meeting*. Filadelfia, 24–28 de enero de 2014.
http://www.bne.es/webdocs/Inicio/Perfiles/Bibliotecarios/ALA/ALAFiladelfiaMidwinter2014.pdf

Véase el apartado 4.1.2: "Harvesting and normalization at the *Digital Public Library of America*: Lessons from a diverse aggregation".

En él se mencionan los vocabularios utilizados para hacer el enriquecimiento, y se dice explícitamente:
"¿Cómo empezar? Es necesario limpiar los datos con herramientas como *OpenRefine*, *Data Wrangler* y *GREL/regez*".

3. Free Your Metadata.
https://goo.gl/BUCaq2

4. Véase por ejemplo:

PCC Task Group on URIs in MARC.
https://www.loc.gov/aba/pcc/bibframe/TaskGroups/URI-TaskGroup.html

También se puede consultar:
Bibframe training at the Library of Congress https://www.loc.gov/catworkshop/bibframe

Y, por supuesto:
Morillo-Calero, María-Jesús (2017). *Impact of linked data project datos.bne.es on authority control at the National Library of Spain*.
https://goo.gl/TzjiJa

5. International Linked Data Survey.
http://www.oclc.org/research/themes/data-science/linkeddata.html#linked-data-survey

6. Entre los numerosos proyectos véase *Linked Data for Production* (*LD4P*):
https://goo.gl/sNvU4m

Y los ya mencionados informes de las reuniones de ALA que elabora María Jesús Morillo Calero, el *punto* 6.4. "Prepararse para ser enlazado: mejorar los datos MARC con URIs con un presupuesto ajustado" en:

Morillo-Calero, María-Jesús (2014). *Informe ALA Midwinter Meeting*. Filadelfia, 24–28 de enero de 2014.
http://www.bne.es/webdocs/Inicio/Perfiles/Bibliotecarios/ALA/ALAFiladelfiaMidwinter2014.pdf

7. *Wikipedia:Authority control integration proposal*
https://en.wikipedia.org/wiki/Wikipedia:Authority_control_integration_proposal

8. *Plantilla:Control de autoridades*
https://es.wikipedia.org/wiki/Plantilla:Control_de_autoridades

9. *Wikidata:WikiProject Authority control*
https://www.wikidata.org/wiki/Wikidata:WikiProject_Authority_control

10. *Wikipedia:Artículos con control de autoridades*
https://goo.gl/H5nkbG

11. *Wikidata:Fifth Birthday.*
https://www.wikidata.org/wiki/Wikidata:Fifth_Birthday

12. Nota a esta edición digital. *Biblioteca Virtual de Viajes Científicos Ilustrados*.
https://goo.gl/sM6DX9

5. Referencias

Abele, Andrejs; **McCrae, John P.**; **Buitelaar, Paul**; **Jentzsch, Anja**; **Cyganiak, Richard** (2017). *Linking open data cloud diagram 2017*.
http://lod-cloud.net

Agenjo, Xavier; Hernández, Francisca; **Viedma, Andrés** (2011). "Data aggregation and dissemination of authority records through linked open data ". En: *IFLA 2011. Meeting: Cataloguing section*.
https://www.ifla.org/past-wlic/2011/80-agenjo-en.pdf

Agenjo, Xavier; Hernández, Francisca; **Viedma, Andrés** (2012). "Data aggregation and dissemination of authority records through linked open data in a European context". *Cataloging & classification quarterly*, v. 50, n. 8.
https://doi.org/10.1080/01639374.2012.711441

Baker, Thomas; **Bermès, Emmanuelle**; **Coyle, Karen**; **Dunsire, Gordon**; **Isaac, Antoine**; **Murray, Peter**; **Panzer, Michael**; **Schneider, Jodi**; **Singer, Ross**; **Summers, Ed**; **Waites, William**; **Young, Jeff**; **Zeng, Marcia** (2011). *Informe final del Grupo Incubador de Datos Vinculados de Bibliotecas*. W3C Incubator Group Report.
http://www.larramendi.es/LAM/Incubator/lld/XGR-lld-20111025.html

Charles, Valentine (2014). "The Polymath Virtual Library and EDM". *Europeana pro*, 4 diciembre.
https://pro.europeana.eu/page/polymath-edm

Erxleben, Fredo; **Günther, Michael**; **Krötzsch, Markus**; **Mendez, Julian**; **Vrandeăić, Denny** (2014). "Introducing Wikidata to the linked data web". En: Mika, Peter; Tudorache, Tania; Bernstein, Abraham; Welty, Chris; Knoblock, Craig; Vrandeăić, Denny; Groth, Paul; Noy, Natasha; Janowicz, Krzysztof; Goble, Carole (eds.). *Proceedings of the 13th International Semantic Web Conference. Part I* (*ISWC '14*), Springer-Verlag New York, Inc., New York, USA, pp. 50-65.
https://doi.org/10.1007/978-3-319-11964-9_4

Isaac, Antoine; **Manguinhas, Hugo**; **Stiller, Juliane**; **Charles, Valentine** (2015). *Task force on enrichment and evaluation*. Europeana, 29 octubre.
https://goo.gl/uTiaHE

Klein, Maximillian; **Kyrios, Alex** (2013). "VIAFbot and the integration of library data on Wikipedia". *Code4Lib journal*, n. 22.
http://journal.code4lib.org/articles/8964

Lih, Andrew; **Fernandez, Robert**. "Wikidata, a rapidly growing global hub, turns five". *Wikimedia*, 30 octubre.
https://blog.wikimedia.org/2017/10/30/wikidata-fifth-birthday

Morillo-Calero, María-Jesús (2014). *Informe ALA Midwinter Meeting*. Filadelfia, 24 – 28 de enero de 2014.
https://goo.gl/nfDTof

Patnab (2015). "5 data cleansing tools". *Data science central*, 14 diciembre.
https://www.datasciencecentral.com/profiles/blogs/5-data-cleansing-tools

Smith-Yoshimura, Karen (2016). "Analysis of international linked data survey for implementers". *D-Lib magazine*, v. 22, n. 7/8.
https://doi.org/10.1045/july2016-smith-yoshimura

Stiller, Juliane; **Isaac, Antoine**; **Petras, Vivien** (2014) *EuropeanaTech task force on a multilingual and semantic enrichment strategy: final report*. Europeana, 7 abril.
https://goo.gl/qn5iMc

Thereaux, Olivier (2006). "Content negotiation: why it is useful, and how to make it work". *W3C Blog*, 21 febrero.
https://www.w3.org/blog/2006/02/content-negotiation

Anexo. Fuentes españolas de identificadores para el control de autoridades en *Wikidata*
http://tinyurl.com/y7lpzbyk

Nombre Identificador	Entidad	Web	Descripción
Biblioteca Virtual de Polígrafos ID	*Fundación Ignacio Larramendi*		Identificador para personas en la *Biblioteca Virtual de Polígrafos*
CPE atleta ID		*http://www.paralimpico.es/publicacion/buscador/biografias.asp*	Perfil de una persona en la web del *Comité Paralímpico Español*
Código Nacional de Ocupaciones 2011		*http://www.ine.es/daco/daco42/clasificaciones/corr_cno11_ciuo08.xls*	
Diccionario Biográfico de Mujeres			
LEMb	*Ministerio de Educación, Cultura y Deporte*	*http://id.sgcb.mcu.es/*	Encabezamientos de materia mantenido por el *Ministerio de Educación, Cultura y Deporte de España*
Código *Bien Cultural de Interés Nacional*			
Código *Bien de Interés Cultural*		*http://www.mecd.gob.es/bienes/cargarFiltroBienesInmuebles.do?layout=bienesInmuebles*	Identificador de un elemento en la base de datos de bienes inmuebles del *Registro de Bienes de Interés Cultural* del *Ministerio de Cultura de España*
Código *IGPCV*		*http://www.ceice.gva.es/web/patrimonio-cultural-y-museos/inventario-general*	Código de identificación de un bien cultural en el *Inventario General del Patrimonio Cultural Valenciano*
Código *IGPCV*		*http://www.cult.gva.es/dgpa/brl/brl.asp*	Código de identificación de un bien cultural en el *Inventario General del Patrimonio Cultural Valenciano*
Código *INE*		*http://www.ine.es/daco/daco42/codmun/codmunmapa.htm*	Código asignado a las entidades de población por el *Instituto Nacional de Estadística de España*
Código *Inventario del Patrimonio Arquitectónico de Cataluña*	*Joan Tuset i Suau*		
Código *SIPCA*		*http://www.sipca.es/censo/busqueda_simple.html*	Identificador de un item en el *Sistema de Información del Patrimonio Cultural Aragonés* (*SIPCA*)
Identificador *AELG*	*Asociación de Escritores en Lengua Gallega*		Identificador de un autor en la página web de la *Asociación de Escritores en Lengua Gallega*
Identificador *As*			Identificador de un deportista en *as.com*
Identificador *Auñamendi*	*Enciclopedia Auñamendi*		Identificador de un elemento en la *Enciclopedia Auñamendi*
Identificador *BDCYL* de autoridad	*Biblioteca Digital de Castilla y León*	*http://bibliotecadigital.jcyl.es/es/consulta/indice_campo.cmd?campo=idautor&letra=A&posicion=1*	Identificador de un autor, tema o lugar en la *Biblioteca Digital de Castilla y León*
Identificador *BNE*	*Biblioteca Nacional de España*		Identificador de la *Biblioteca Nacional de España*
Identificador *BNE de publicación periódica*	*Hemeroteca Digital* de *la Biblioteca Nacional de España*	*http://hemerotecadigital.bne.es/index.vm*	Identificador de un periódico o revista en la *Hemeroteca Digital* de la *Biblioteca Nacional de España*
Identificador *BVMC de autor*	*Biblioteca Virtual Miguel de Cervantes*	*https://data.cervantesvirtual.com*	Identificador de un autor en la *Biblioteca Virtual Miguel de Cervantes*
Identificador *BVMC de obra*	*Biblioteca Virtual Miguel de Cervantes*	*http://data.cervantesvirtual.com/*	Identificador de una obra en la *Biblioteca Virtual Miguel de Cervantes*
Identificador *BVPH*	*Biblioteca Virtual de Prensa Histórica*	*http://prensahistorica.mcu.es*	Identificador en la *Biblioteca Virtual de Prensa Histórica*

Identificador *Biblioteca Valenciana Digital de autor*		*http://bivaldi.gva.es/en/cms/elemento.cmd?id=estaticos%2Fpaginas%2Finicio.html*	Identificador de un autor en la *Biblioteca Valenciana Digital* (*BiValDi*)
Identificador *CANTIC*		*http://www.bnc.cat*	Catálogo de autoridades de nombres y títulos de Cataluña, gestionado por la *Biblioteca Nacional de Cataluña*
Identificador *COAM inmueble*			Identificador de un edificio o estructura destacados en la ciudad de Madrid en la base de datos del *Colegio Oficial de Arquitectos de Madrid* (*COAM*)
Identificador *COAM persona*		*http://212.145.146.10/biblioteca/fondos/ingra2014/index.htm#car.webA*	Identificador de un individuo que ha construido o proyectado edificios o estructuras significativas en la ciudad de Madrid disponibles en la base de datos del *Colegio Oficial de Arquitectos de Madrid* (*COAM*)
Identificador *DBSE*	*Diccionario biográfico del socialismo español*	*http://www.fpabloiglesias.es/archivo-y-biblioteca/diccionario-biografico*	Identificador de una persona en el Diccionario biográfico del socialismo español
Identificador *DBSE*	*Fundación Pablo Iglesias*	*http://www.fpabloiglesias.es/archivo-y-biblioteca/diccionario-biografico*	Identificador de una persona en el *Diccionario biográfico del socialismo español*
Identificador *DOCOMOMO Ibérico*	*Fundación Docomomo Ibérico*	*http://www.docomomoiberico.com/index.php?option=com_content&view=article&id=43&Itemid=61*	Identificador de un edificio, estructura o conjunto de edificios en la base de datos *Registros del Movimiento Moderno*, de *DOCOMOMO Ibérico*
Identificador *Diccionario biográfico español*		*http://www.rah.es:8888*	Identificador de una persona con artículo en el *Diccionario biográfico español*
Identificador *Eldoblaje* de película		*http://www.eldoblaje.com/*	Identificador de doblaje en la web *eldoblaje.com*, una base de datos de doblajes españoles
Identificador *FilmAffinity*			Número de identificación *Filmaffinity*
Identificador *Galiciana* de autor	*Galiciana*	*http://galiciana.bibliotecadegalicia.xunta.es/i18n/estaticos/contenido.cmd?pagina=estaticos/presentacion*	Identificador de un autor en *Galiciana*
Identificador *Galiciana* de obra	*Galiciana*	*http://galiciana.bibliotecadegalicia.xunta.es/i18n/estaticos/contenido.cmd?pagina=estaticos/presentacion*	Identificador de una obra en *Galiciana*
Identificador *Gran Enciclopedia Aragonesa*			Identificador de una persona con artículo en *Gran Enciclopedia Aragonesa*
Identificador *MNCARS artista*		*http://www.museoreinasofia.es/en/authors*	Identificador de artista en el *Museo Nacional Centro de Arte Reina Sofía* (*MNCARS*)
Identificador *Patrimonio Inmueble de Andalucía*	*Instituto Andaluz del Patrimonio Histórico*	*http://www.iaph.es/patrimonio-inmueble-andalucia/frmSimple.do*	Identificador de un bien cultural en la base de datos *Patrimonio Inmueble de Andalucía* (*BDI*)
Identificador *Patrimonio Web JCyL*		*http://servicios.jcyl.es/pweb/buscarInmueble.do*	Identificador de un ítem en *Patrimonio Web* de la *Junta de Castilla y León*
Identificador *RANM*		*http://www.ranm.es/academicos.html*	Identificador de un miembro de la *Real Academia Nacional de Medicina* (*RANM*) de España
Identificador *SNCZI-IPE de embalse*		*http://sig.mapama.es/snczi/visor.html*	Identificador de un embalse en España, en *SNCZI-Inventario de Presas y Embalses*
Identificador *SNCZI-IPE de presa*		*http://sig.mapama.es/snczi/visor.html*	Identificador de una presa en España, en *SNCZI-Inventario de Presas y Embalses*
Identificador *Thyssen-Bornemisza de artista*	*Museo Thyssen-Bornemisza*	*http://www.museothyssen.org/en/thyssen/artistas*	
Identificador *Universidad de Barcelona*			Identificador externo

Identificador *acb.com*		*http://www.acb.com*	Identificador de un jugador de baloncesto en *acb.com*
Identificador de la *Gran Enciclopedia Catalana*	*Gran Enciclopedia Catalana*	*https://www.enciclopedia.cat*	Identificador de la *Gran Enciclopedia Catalana*
Identificador de monumento de Zaragoza			Identificador de un monumento catalogado en *zaragoza.es*
Identificador del *Boletín Oficial del Estado*			Identificador de un decreto, ley, orden o en general cualquier anuncio oficial en el *Boletín Oficial del Estado*, publicación oficial del Gobierno de España

Xavier Agenjo Bullón
Fundación Ignacio Larramendi
xavier.agenjo@larramendi.es

Francisca Hernández Carrascal
DIGIBÍS, Producciones Digitales
francisca.hernandez@digibis.com