Análisis de la anotación temática de los recursos culturales en la Web de Datos

Dayany Díaz-Corona, Javier Lacasta, Javier Nogueras-Iso

Grupo de Sistemas de Información Avanzados (IAAA)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: dayanydc@unizar.es

Resumen

Este trabajo presenta una metodología para el estudio de modelos de organización del conocimiento (vocabularios controlados, tesauros, ontologías) utilizados en la anotación temática de recursos de patrimonio cultural en la web de datos: desde la identificación de las fuentes de datos, hasta la selección, acceso y análisis de la calidad de los modelos de conocimiento.

Introducción

Gracias a políticas de preservación de patrimonio cultural hoy en día es posible acceder a través de la web a abundante material (documentos, fotografías, grabaciones) que se ha digitalizado y publicado junto con metadatos en repositorios digitales. En el dominio del patrimonio cultural, el carácter enlazado de la información ha promovido la aplicación de tecnologías de la web semántica para la construcción de estos repositorios y publicarlos como Datos Enlazados (*Linked Data*) [1], una iniciativa promovida por W3C para describir formalmente los recursos con RDF y definir hiperenlaces entre recursos de la Web creando así una Web de Datos.

Un aspecto esencial de cualquier repositorio semántico es la calidad y precisión de su anotación (metadatos). De todas las propiedades posibles para la anotación de un recurso, las que describen la temática son las más importantes para facilitar su descubrimiento. Por tanto, la categoría o tema de los recursos se suele seleccionar de modelos de organización del conocimiento predefinidos. Sin embargo, si la estructura y calidad de estos modelos no es apropiada, se reduce drásticamente su utilidad. Este trabajo hace un estudio de los modelos de organización del conocimiento utilizados en la web de datos para recursos culturales.

Metodología

El proceso propuesto para analizar la calidad y la idoneidad de los modelos de conocimiento utilizados en la web de datos de recursos culturales consta de los siguientes pasos:

1) Adquisición de datos: identificar fuentes de datos enlazados relevantes para el dominio deseado, así como los métodos de acceso y los modelos de representación utilizados en estas fuentes.
2) Identificación de los modelos de conocimiento utilizados: estudiar el modelo de representación de cada fuente para identificar las propiedades de anotación temática que enlazan a modelos de conocimiento, y posteriormente identificar los distintos espacios de nombres que caracterizan cada modelo de conocimiento.
3) Selección y adquisición de los modelos de conocimiento: filtrar y acceder a aquellos modelos de conocimiento temáticos utilizados por un número relevante de organizaciones y disponibles en un formato procesable.
4) Análisis de la calidad: calcular métricas de calidad sobre de los modelos de conocimiento adquiridos en el paso previo. Este análisis se basa en las 14 métricas propuestas en un trabajo previo [2], que aquí se agrupan en relación a su capacidad para analizar la completitud (existencia de propiedades obligatorias como etiquetas preferidas), consistencia (chequeo del contenido de etiquetas para asegurar, por ejemplo, un uso uniforme de singulares/plurales o mayúsculas) y exactitud (entre otros aspectos, chequeo de la corrección y exactitud de relaciones jerárquicas entre conceptos).

Resultados

Mediante búsquedas en la web y revisión de estudios de fuentes de Datos Enlazados, se han identificado las siguientes fuentes de datos: la biblioteca digital *Europeana* (54,9 millones de recursos); la *Digital Public Library of America (DPLA)* (20,7 millones); la *British National Bibliography* (2,8 millones); la *Public Library of Veroia* (3,1 millones); y el *Repository for Linked Open Archival Data* (0,18 millones).

En todos los casos, el contenido de estos repositorios se describe utilizando RDF y aunque reutilizan propiedades de modelos como Dublin Core y FOAF, cada uno tiene propiedades diferentes. Respecto al acceso, salvo DPLA, todos

los repositorios facilitan un SPARQL *end-point* para acceder al contenido.

Analizando los proveedores de datos y los modelos de conocimiento utilizados para la anotación, hemos identificado 433 proveedores de datos diferentes que clasifican sus recursos con 49 millones de palabras clave. Sin embargo, solo la mitad (26 millones) son referencias a modelos de conocimiento (URLs), siendo el resto texto libre.

Las URLs extraídas de los repositorios se han procesado para identificar los espacios de nombres de los modelos de conocimiento. Hemos encontrado 58 modelos de conocimiento que incluyen tesauros, taxonomías y listas controladas. La Fig. 1 (izquierda) muestra las estadísticas sobre el tipo de contenido: 47 son modelos temáticos; 6 son archivos de autoridad; 2 de ellos son nomenclátores; y 3 de ellos son combinaciones de diferentes tipos. Con respecto al formato (Fig. 1, arriba a la derecha), 32 siguen el modelo SKOS [3], 16 utilizan sus propios esquemas RDF (13) o OWL (3), y otros 10 están en formatos no semánticos (HTML o texto).

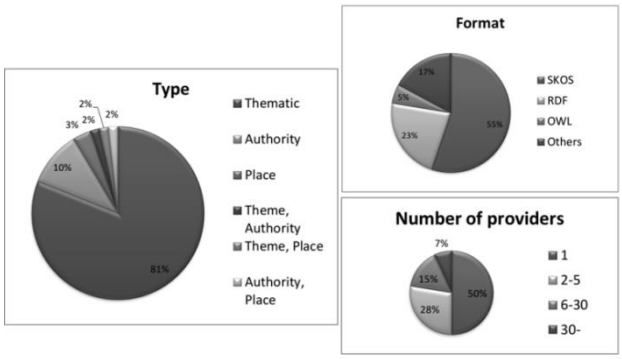

Figura 1: Características de los modelos de conocimiento

Además, hemos analizado cómo los proveedores de datos utilizan estos modelos para identificar los más relevantes (Fig. 1, abajo a la derecha). La mitad de ellos (29) han sido utilizados por un único proveedor. Una revisión detallada ha demostrado que la mayoría de estos son modelos ad-hoc definidos por una sola organización para clasificar sus propios recursos, lo que contradice el propósito de la web semántica de reutilizar y extender modelos. Otros 16 modelos han sido empleados por 2-5 proveedores. Son algo similares a los del primer caso, ya que una revisión ha demostrado que estos pequeños grupos de proveedores están de alguna manera relacionados entre ellos. Hay 9 modelos que han sido utilizados por 6-30 proveedores Su revisión ha demostrado que hay comunidades de usuarios que contribuyen a su diseño, mejorando su estructura y forma de distribución. Finalmente, hay 4 modelos que se utilizan por más de 50 proveedores y representan alrededor del 60% de los usos del modelo, lo que muestra la concentración del campo en torno a unos pocos modelos. Esto es muy relevante ya que demuestra que es posible distinguir las partes de la web semántica que están mejor conectadas y descritas.

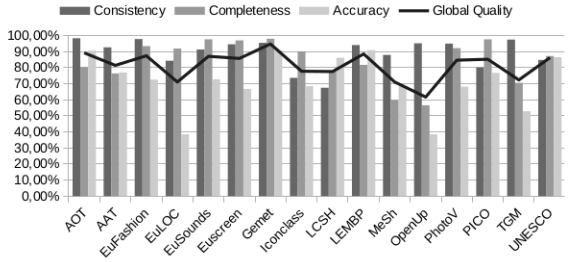

Figura 2: Calidad de los modelos de conocimiento

Por último, la Fig. 2 muestra los resultados del análisis de la calidad de 16 modelos de conocimiento temáticos utilizados por más de un proveedor. Aunque en general el promedio de la completitud, consistencia y exactitud es alto, algunos modelos como OpenUp (vocabulario de historia natural) presenta importantes problemas de completitud.

Conclusiones

Este trabajo ha mostrado el análisis realizado al subconjunto de la web de datos relacionado con los recursos del patrimonio cultural para identificar los modelos de conocimiento utilizados en la clasificación de sus recursos y valorar su calidad.

Además, este estudio nos ha permitido descubrir varios problemas que no son solo del dominio del patrimonio cultural, sino en general de la web semántica: no hay herramientas de búsqueda específicas para localizar fuentes de datos enlazados; hay un problema de heterogeneidad de modelos, incluso para propiedades básicas como el tema; y falta semántica en los valores de las propiedades, que muchas veces son literales en lugar de URIs referenciando correctamente a otros recursos.

AGRADECIMIENTOS

Este trabajo ha sido financiado parcialmente por el Gobierno de España a través del proyecto TIN2017-88002-R. El trabajo de Dayany Díaz-Corona ha sido financiado por una beca de la Universidad de Zaragoza y el Banco Santander.

REFERENCIAS

[1] BERNERS-LEE, T. Linked Data Web architecture note, 27 July 2006. http://www.w3.org/DesignIssues/LinkedData.html

[2] LACASTA, J., FALQUET, G., ZARAZAGA-SORIA, F.J., and NOGUERAS-ISO, J.. An automatic method for reporting the quality of thesauri. *Data & Knowledge Engineering*. 2016, 104, 1-14.

[3] MILES, A., and BRICKLEY, D. SKOS Core Guide, 2005. http://www.w3.org/TR/swbp-skos-core-guide/