Enriquecimiento automático de portales culturales mediante modelos de organización del conocimiento

Automatic enrichment of cultural portals with knowledge organization systems

Dayany DÍAZ-CORONA (1), Javier LACASTA (2), Javier NOGUERAS-ISO (3)

(1) Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, C/ María de Luna, 1, 50018 Zaragoza, España, dayanydc@unizar.es (2) jlacasta@unizar.es (3) jnog@unizar.es

Resumen

Durante las últimas décadas se han ido creando numerosos portales web para diseminar el patrimonio cultural. La mayoría de estos portales se crearon en tiempos de la web sintáctica generando páginas HTML con texto plano indexable por buscadores, pero sin metadatos añadidos y anotaciones de conceptos pertenecientes a modelos de organización del conocimiento que facilitarían la labor de buscadores temáticos especializados. Este artículo propone un método para recomendar los modelos de organización del conocimiento que mejor se ajusten a los contenidos de un portal web, y utilizar esos modelos para anotar semánticamente los contenidos. Para verificar la viabilidad del método propuesto se ha aplicado en el enriquecimiento de un portal creado a mediados de los años 90 y que aloja un catálogo virtual de las obras del pintor Goya. Gracias al método propuesto, se ha recomendado el modelo de organización del conocimiento denominado Lista de Encabezamientos de Materias para las Bibliotecas Públicas por su cercanía con el contenido del portal. Además, se han conseguido anotar semánticamente dos tercios de las páginas en castellano del portal con conceptos de este modelo. Aunque la exactitud de los emparejamientos entre las entidades detectadas en el texto y los conceptos del modelo no es perfecta, la anotación realizada constituye una buena base para que los administradores del portal puedan refinar posteriormente esta anotación.

Palabras clave: Modelos de organización del conocimiento. Sistemas de organización del conocimiento. SKOS. Web semántica. Metadatos. Sistemas de recomendación.

Abstract

During the last decades, numerous web portals have been launched to disseminate the cultural heritage. Most of these portals were developed with technologies from the syntactic web era, i.e. containing HTML pages with plain text that can be indexed by search engines, but without additional metadata and annotations of concepts belonging to knowledge organization systems that would facilitate the task of thematic specialized search engines. This paper proposes a method for recommending the knowledge organization systems that are better adjusted for the contents of a web portal and the use of these systems for the semantic annotation of the contents. To check the feasibility of the proposed method, we have applied it to the enrichment of a web portal created in the nineties that hosts a virtual catalogue of the works performed by the painter Goya. Thanks to the proposed method, we have been able to recommend knowledge organization system titled List of Subject Headings for Public Libraries because of its closeness with the portal content. In addition, two thirds of the web pages in Spanish were annotated with concepts belonging to this model. Although the accuracy of the mapping between the recognized entities in the text and the concepts of the model is not perfect, it constitutes a good base to allow web portal administrators to refine later this annotation.

Keywords: Knowledge organization systems. SKOS. Semantic web. Metadata. Recommendation systems.

1. Introducción

Desde mediados de los años 90 se ha fomentado la digitalización del patrimonio cultural creándose numerosos portales web para difundir ese patrimonio a todos los públicos. La mayoría de estos portales se crearon con tecnologías pertenecientes a la época de la web sintáctica y contienen páginas HTML con texto plano indexable por buscadores. Pero para que estos portales sean indexables por motores de búsqueda y agregadores especializados y que sus contenidos sean fácilmente accesibles por el público, es esencial anotar semánticamente sus contenidos con palabras clave temáticas extraídas de modelos de organización del conocimiento de alta calidad. Estos modelos de organización del conocimiento son vocabularios especializados que incluyen tesauros y cuadros de clasificación, entre otros modelos, y que al utilizarlos de forma consensuada minimizan lo máximo posible los problemas derivados de la heterogeneidad del lenguaje como la

polisemia o la sinonimia (Lacasta et al., 2010; Whaley et al, 2020).

En los últimos años han surgido distintas aproximaciones para analizar la calidad de modelos de organización del conocimiento (Suominen y Mader, 2014; Albertoni et al., 2016; Lacasta et al., 2016). Este análisis de calidad permite que los creadores de contenido puedan seleccionar un modelo en función de su calidad. Sin embargo, este análisis no tiene en cuenta si existe un emparejamiento adecuado entre los contenidos que se quieren anotar temáticamente y los conceptos cubiertos por el modelo de organización del conocimiento. Algunos autores como da Silva Lemos y Souza (2020) proponen una serie de criterios para poder comparar las ventajas y desventajas del uso de distintos sistemas de organización del conocimiento, pero no es un procedimiento automatizado.

El objetivo de esta contribución es proponer un método que permita recomendar de forma automática los modelos de organización del conocimiento que mejor cubran los contenidos de un portal web que se quieren enriquecer semánticamente y facilitar así la capacidad de búsqueda sobre esos contenidos. Es decir, partiendo de un conjunto de modelos de organización del conocimiento con calidad contrastable, se pretende elegir aquel que sea más adecuado al contenido que se quiere anotar.

En la literatura existen varias propuestas de anotadores semánticos de contenido textual de carácter general como DBPedia Spotlight (Daiber et al., 2013) o RDFFace (Khalili y Auer, 2015). Estas herramientas permiten detectar secuencias de palabras correspondientes a entidades nombradas (o simplemente entidades) que hacen referencia a materias, lugares u organizaciones y las enlazan con conceptos de ontologías o grafos de conocimiento como DBPedia (en el caso de DBPedia Spotlight) u otras fuentes consultables a través de APIs como Swoogle o Sindice. También han surgido propuestas similares de anotación en contextos más específicos como el de la medicina (Thessen y Parr, 2014; Tchechmedjiev et al., 2018) o el sector multimedia (Rodriguez Rocha et al, 2015) focalizadas en el uso de ontologías más específicas o añadiendo técnicas más expertas en el reconocimiento de entidades en estos contextos.

A diferencia de las soluciones existentes, la novedad del trabajo que se propone en este artículo es establecer un flujo de trabajo que permita flexibilizar las ontologías elegibles para la anotación. El método propuesto en este artículo parte de un trabajo previo de los autores (Díaz-Corona et al., 2019) donde se estableció un proceso para caracterizar los modelos de organización del conocimiento más utilizados en el dominio cultural y que además incluye un análisis de la completitud, consistencia y exactitud de los modelos compatibles con SKOS (Miles y Brickley, 2005), una iniciativa de W3C para representar modelos de organización del conocimiento utilizando un vocabulario RDF. A continuación, para identificar si los modelos con una calidad razonable cubren el contenido que se quiere anotar se evalúa su similitud en el espacio vectorial. Y finalmente, para anotar el contenido con conceptos concretos el método se apoya en software existente de reconocimiento de entidades.

Como banco de pruebas para estudiar la viabilidad de este método de recomendación se ha utilizado el portal InfoGoya (Universidad de Zaragoza, 1996), un catálogo virtual de las obras del pintor Goya creado a mediados de los 90. Es un ejemplo de un portal de la web sintáctica que se puede enriquecer con micro-metadatos para facilitar su semantización y mejorar su indexación.

El resto del artículo se estructura de la siguiente forma. En la sección 2 se explica el método propuesto. En la sección 3 se muestra el resultado de aplicar el método sobre el portal mencionado anteriormente. Finalmente, la sección 4 presenta las conclusiones finales y las líneas de trabajo futuro.

2. Método

La figura 1 muestra el procedimiento propuesto en este trabajo para el enriquecimiento semántico de portales.

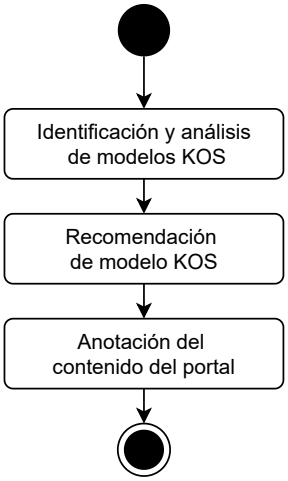

Figura 1. Flujo de trabajo para el enriquecimiento de portales semánticos

El primer paso es la identificación y análisis de potenciales modelos de organización del conocimiento, denominados a partir de este momento

por su acrónimo en inglés KOS (*Knowledge Organization Systems*). Partiendo de la selección previa de modelos KOS, el segundo paso está orientado a recomendar el modelo KOS que resulte más apropiado. El tercer paso es anotar el contenido del portal con el modelo KOS recomendado. En las siguientes secciones se explican los pasos de este procedimiento.

2.1. Identificación y análisis de modelos de organización del conocimiento

Como se ha mencionado en la introducción, para la identificación y análisis de los modelos KOS se parte de un trabajo previo de los mismos autores de este artículo que analizaba los modelos KOS utilizados en repositorios de datos enlazados del dominio cultural (Díaz-Corona et al., 2019). Las fases contempladas en el proceso planteado en este trabajo previo se muestran en la Figura 2.

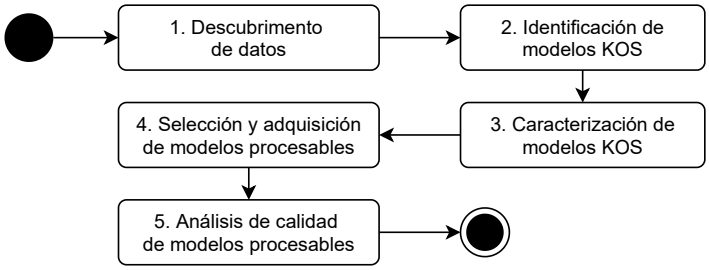

Figura 2. Fases de la metodología para la identificación y análisis de modelos de organización del conocimiento

La primera fase consiste en identificar las fuentes de datos enlazados relevantes para el dominio deseado, así como los métodos de acceso. Es un proceso principalmente manual porque no hay un catálogo único centralizado de fuentes de datos enlazados o un único sistema de búsqueda apropiado para buscar fuentes de datos. Por tanto, es necesario revisar tanto posibles estados del arte referenciando a fuentes de datos como utilizar buscadores de datos (ej: Google Dataset), o incluso buscadores genéricos.

```
<ore:Proxy rdf:about="http://data.europeana.eu/proxy/provider/2059210/data_sounds_http__
            _imslp_org_wiki_Fa__la_nonna_ninno_mio__Sarria__Enrico_">
   <dc:creator rdf:resource="http://imslp.org/wiki/Category:Sarria,_Enrico"></dc:creator>
   <dc:title>Fa' la nonna ninno mio (Sarria, Enrico)</dc:title>
   <dc:subject rdf:resource="http://data.europeana.eu/concept/soundgenres/Music/
            Western_classical_music"></dc:subject>
   <dc:subject xml:lang="en">Romantic</dc:subject>
   <ore:proxyIn rdf:resource="http://data.europeana.eu/aggregation/provider/2059210/
            data_sounds_http___imslp_org_wiki_Fa__la_nonna_ninno_mio__Sarria__Enrico_"/>
</ore:Proxy>
```

Figura 3. Fragmento de los metadatos de un recurso cultural enlazando a un KOS de Europeana

La segunda fase consiste en estudiar el modelo de metadatos de cada fuente con el objetivo de identificar las propiedades de anotación temática que enlazan a conceptos pertenecientes a modelos KOS y extraer un listado de todos los modelos KOS referenciados. Por ejemplo, en la Figura 3 de muestra un fragmento de los metadatos de un recurso sonoro anotado con un concepto del tesauro *Europeana Sounds* a través de la etiqueta *dc:subject*.

La tercera fase es el estudio y caracterización de los modelos identificados en la segunda fase. Como resultado de esta fase se anota cada modelo KOS con los siguientes elementos, entre otros descriptores: información descriptiva como el nombre, su acrónimo y editor; información sobre su relevancia mediante su nº de usos y nº de proveedores de recursos culturales referenciando a estos modelos; tipo de clasificación (temática, de autoridad o de lugar) para la que se utiliza el KOS; y los detalles sobre el mecanismo de acceso y formato de descarga (por ejemplo, SKOS, RDF, OWL, etc.).

La cuarta fase consiste en filtrar y acceder a aquellos modelos KOS utilizados para clasificaciones temáticas de recurso, que hayan sido utilizados por un número relevante de proveedores (dos o más) y estén disponibles en un formato compatible con SKOS.

La última fase está dedicada a analizar en detalle la calidad de los modelos KOS procesables que se descargaron en la fase previa, y que son susceptibles de ser reutilizables para el enriquecimiento semántico de nuevos contenidos. Este análisis, propuesto por Lacasta et al. (2016), está basado en la norma internacional ISO 25964 para tesauros y plantea 14 métricas agrupadas en 3 categorías: consistencia, planteando métricas que chequean el contenido de las etiquetas de los conceptos incluidos en el KOS para asegurar, por ejemplo, un uso uniforme de singulares/plurales o mayúsculas; completitud, agrupando métricas que verifican la existencia de propiedades obligatorias como etiquetas preferidas; y exactitud, incluyendo métricas que chequean, entre otros aspectos, la corrección y exactitud de relaciones jerárquicas entre conceptos.

2.2. Recomendación del modelo KOS

El paso anterior del procedimiento ha conseguido identificar modelos KOS con calidad contrastable y un formato apto (compatibilidad con SKOS) para anotar semánticamente contenidos, pero no hemos verificado si estos vocabularios son adecuados temáticamente para nuestros contenidos.

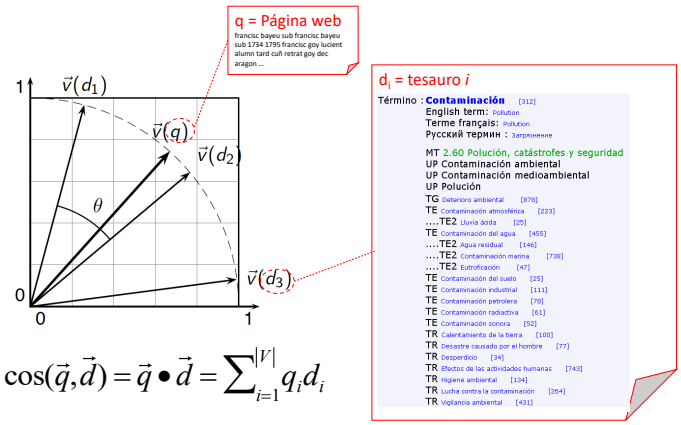

Figura 4. Adaptación del modelo vectorial para el ranking de modelos KOS

Algoritmo: Recomendacion

```
Input: modelosKOS, paginasPortal, idiomaPortal
Output: rankingKOS
 # Transformar sitio web en una colección de documentos
1:  textosPortal = {}
2:  for paginaHTML in paginasPortal do
3:    textoPagina = transformacionATextoPlano(paginaHTML)
4:    textosPortal = add(textosPortal, textoPagina)
5:  end for
 # Transformar modelos KOS en una colección de documentos
6:  textosKOS = {}
7:  for modelo in modelosKOS do
8:    if modelo contiene etiquetas en idiomaPortal then
9:      textoKOS = ""
10:     for concepto in modelo do
11:     etiquetas = {'prefLabel', 'altLabel', 'definition', 'scopeNote'}
12:     for etiqueta in etiquetas do
13:       textoKOS = append(textoKOS,
                  extraer(modelo, concepto, etiqueta, idiomaPortal))
14:     end for
15:     textosKOS = add(textosKOS, <textoKOS, modelo>)
16:   end for
17:   N = length(textosKOS)
18:  end if
19: end for
  #Inicialización de puntuaciones de cada modelo KOS
20: rankingKOS = {}
21: for elemento in textosKOS do
22:   modelo = elemento.getValor()
23:   rankingKOS.put(modelo,0)
24: end for
25: indiceModelosKOS = indexar(textosKOS)
  #Votación de cada página a los modelos KOS
26: for textoPagina in textosPortal do
27:   consulta = parsear(textoPagina)
28:   resultados = buscarModeloVectorial(consulta, indiceModelos-
KOS)
29:   puntuacion = N
30:   for modelo in resultados do
31:     rankingKOS.put(modelo, rankingKOS.get(modelo)+ puntua-
cion)
32:     puntuacion = puntuacion -1
33:   end for
34: end for
  #Ordenar en sentido decreciente los modelos según puntuación
35: rankingKOS = sort_descending(rankingKOS)
36: return rankingKOS
```

Figura 5. Algoritmo de recomendación de modelos KOS

En este paso del procedimiento se pretende establecer un ranking de los modelos KOS más cercanos al contenido de las páginas web del portal que queremos anotar. Para ello se propone adaptar el modelo vectorial de recuperación de información que proporciona un ranking de resultados según la cercanía entre los vectores que representan a los documentos y el vector que representa a la consulta. Utilizando el coseno del ángulo formado entre el vector documento y el vector consulta se devuelve en primer lugar los vectores documento que forman un ángulo cercano a cero respecto al vector consulta. Adaptando el modelo vectorial a nuestro contexto, tal como se indica en la Figura 4 en nuestro caso cada modelo KOS es un documento y las páginas individuales actúan como consultas del sistema de recuperación.

La Figura 5 muestra el algoritmo seguido para establecer un ranking de los modelos KOS recomendados para el contenido del portal. En primer lugar, se procesa cada una de las páginas del portal para transformarlas a texto plano y convertir así el sitio web en una colección de documentos (líneas 1-5). A continuación, se hace un filtrado de los modelos KOS disponibles en el idioma del portal y aquellos que pasan el filtro se transforman también en un documento de texto (líneas 6-19). Para esta transformación a un documento de texto se extraen las etiquetas preferidas (*prefLabel*), las etiquetas alternativas (*altLabel*), las definiciones (*definition*) y las notas de alcance (*scopeNote*). Por último, se establece un sistema de votación de los modelos KOS (líneas 20-36). Para cada página del portal se calculan los modelos KOS más cercanos según el modelo vectorial indicado anteriormente y se otorga una puntuación a los modelos KOS. Si hay N modelos KOS para puntuar, el KOS más cercano a la página recibe N puntos, el siguiente KOS recibe N-1 puntos y así sucesivamente.

Finalmente, el algoritmo devuelve un listado ordenado de los modelos KOS según el sumatorio de las puntuaciones otorgadas por cada página. Respecto al procesamiento de texto utilizado en las líneas 25 y 27 para representar los modelos KOS y las páginas del portal como vectores de palabras, cabe destacar que se ha utilizado un analizador estándar para castellano proporcionado por el motor de indexación Lucene a través de la clase *SpanishAnalyzer*, la cual realiza una conversión a minúsculas de las palabras contenidas en el flujo de entrada de texto, separa los *tokens* con caracteres que no se corresponden con letras, elimina un conjunto básico de palabras vacías en castellano, y extrae las raíces de las palabras con un algoritmo de *stemming* (Savoy, 2002).

2.3. Anotación del contenido del portal.

Una vez que ya sabemos cuál es el modelo KOS más adecuado para anotar el contenido del portal, el último paso es procesar el texto de cada página e incluir los metadatos con los conceptos del modelo KOS que mejor se ajustan a ese texto.

Algoritmo: Anotacion

```
Input: modeloKOS, paginasPortal
Output: paginasAnotadas
1:  paginasAnotadas = {}
2:  grafoRDF = cargarGrafo(modeloKOS)
3:  for pagina in paginasPortal do
4:    tiposEntidad = {'subject', 'location', 'organization', …}
5:    anotaciones = {}
6:   for tipoEntidad in tiposEntidad do
7:     entidades = extraerEntidades(pagina, tipoEntidad)
8:     for entidad in entidades do
9:       etiquetas = {'prefLabel', 'altLabel', 'definition', 'scopeNote'}
10:      concepto = grafoRDF.query(entidad, tipoEntidad, etiquetas)
11:      end for
12:      anotaciones = add(anotaciones, concepto)
13:   end for
14:  paginasAnotadas = add(<pagina, anotaciones>)
15:  end for
16:  return paginasAnotadas
```

Figura 6. Algoritmo de anotación

La Figura 6 indica el procedimiento seguido para realizar la anotación de cada página. En primer lugar, se carga el grafo RDF correspondiente al modelo KOS seleccionado (línea 2) para poder realizar consultas sobre él utilizando el lenguaje SPARQL. Posteriormente, para cada página, se extraen las expresiones correspondientes a entidades nombradas, o simplemente entidades, que hacen referencia a materias, lugares u organizaciones (línea 7) con una librería de procesamiento de lenguaje natural como Apache OpenNLP (Apache, 2017). A continuación, utilizando esas expresiones se buscan conceptos del modelo KOS que tengan algún emparejamiento textual en sus etiquetas *prefLabel*, *altLabel*, *definition* o *scopeNote* utilizando SPARQL en combinación con el motor de indexación Lucene. Además, se tiene en cuenta que el modelo KOS haya podido definir esquemas de conceptos separados para cada tipo de entidad (materia, lugar, etc.).

3. Experimentos iniciales

Tal como se comentó en la introducción, el método propuesto en este trabajo se ha aplicado al portal InfoGoya (Universidad de Zaragoza, 1996), un catálogo virtual de las obras del pintor Goya. Este portal consta de 2.398 páginas HTML: 1.576 páginas en castellano y 822 páginas con la traducción al inglés de los elementos principales del portal. Para el propósito de este experimento nos hemos centrado exclusivamente en el uso de las páginas disponibles en castellano.

Para el primer paso de identificación y análisis de modelos KOS, se partió directamente de los resultados obtenidos por Díaz-Corona et al. (2019) donde se analizaron 5 repositorios de datos enlazados de relevancia en el dominio cultural, entre ellos Europeana (Comisión Europea, 2021) y la Digital Public Library of America (DPLA, 2021), y como resultado final se identificaron 16 modelos KOS con formatos compatibles con SKOS utilizados por un gran número de proveedores.

En la Tabla I se muestran los 8 modelos KOS disponibles en castellano con detalles sobre el número de conceptos en cada idioma (por ejemplo, columna *PL(es)* para castellano), relaciones BT-NT (*broader term – narrower term*), relaciones RT y la valoración global de su calidad. Como se puede observar, los modelos KOS se muestran ordenados de mayor a menor calidad. En general, la calidad global de la construcción de los modelos KOS es alta: 7 de los 8 modelos superan el 80% y el modelo mejor valorado es el tesauro GEMET, un tesauro propuesto en el dominio de medio-ambiente pero que también incluye términos generales en otros dominios. Aunque no es propiamente un tesauro cultural, este tesauro se utiliza en Europeana para interrelacionar recursos de clasificados con distintos modelos KOS.

Después de aplicar el sistema de votación descrito en la sección 2.2 (algoritmo de la Figura 5) a los modelos KOS candidatos, se obtuvo el ranking de puntuaciones que se muestra en la Figura 7. La puntuación de cada modelo KOS también se muestra en la columna *Puntuación* de la Tabla I.

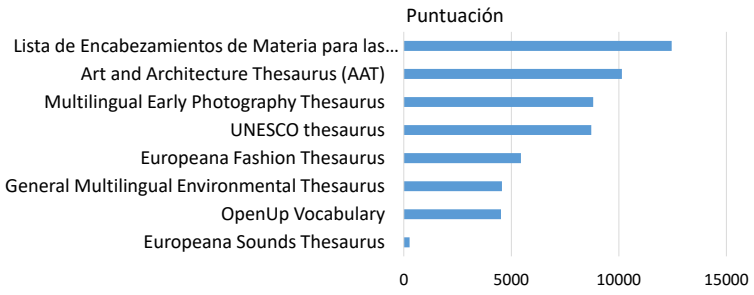

Figura 7. Ranking de modelos KOS

Como se puede observar, el modelo ganador es la Lista de Encabezamientos de Materias para las Bibliotecas Públicas (LEMb). Este modelo ha sido elaborado por la Subdirección General de Coordinación de Bibliotecas, y consta de 19.152 concep-

tos con etiquetas preferidas, etiquetas alternativas y notas de alcance en castellano. El ranking de modelos KOS difiere completamente de la ordenación por calidad mostrada en la Tabla I. Nuestro sistema de recomendación ha conseguido colocar a los modelos sobre cultura y arte en las primeras posiciones y desplazar a los modelos del dominio de las ciencias naturales como GEMET y Open Up a las últimas posiciones.

Modelo KOS	*#*	*PL(en)*	*PL(es)*	*PL(fr)*	*BT-NT*	*RT*	*Calidad*	*Puntuación*
General Multilingual Environmental Thesaurus (Gemet)	5.244	5.235	5.236	5.235	6.555	2.086	94.59%	4.566
Lista de Encabezamientos de Materia para las Bibliotecas Públicas (LEMb)	19.152	0	19.142	0	7.227	24.875	88.48%	12.456
Europeana Fashion Thesaurus	1.089	1.088	799	1.088	1.147	0	87.34%	5.450
Europeana Sounds Thesaurus	15	15	15	15	9	0	87.00%	268
UNESCO Thesaurus	4.424	2.952	3.431	3.295	3.192	12.015	86.16%	8.718
Multilingual Early Photography Thesaurus	553	553	553	553	0	671	84.55%	8.803
Art and Architecture Thesaurus (AAT)	45.942	42.508	31.078	3.548	369.170	18.165	81.28%	10.142
OpenUp Vocabulary	733.097	102.077	19.640	1.195	0	0	61.56%	4.523

Tabla I. Modelos KOS candidatos para la anotación del portal

Por último, se ha seguido el procedimiento propuesto para la anotación (algoritmo incluido en la Figura 6, y explicado en la sección 2.3) con el objetivo de enriquecer las páginas en castellano del portal InfoGoya con conceptos del modelo LEMb. Para cada página se han extraído entidades nombradas de tipo materia y de tipo lugar dado que los conceptos de LEMb se distribuyen en dos esquemas separados: 'Encabezamientos de materia' y 'Encabezamientos de lugar'. De las 1.576 páginas en castellano, se han anotado de forma automática 1.075 páginas (68,21% del total en castellano) con una media de 2,34 conceptos por página. En la Tabla II se muestra el ejemplo de la anotación de una página que contiene el contenido de una carta enviada por el pintor en castellano antiguo donde se han reconocido dos entidades de lugar (*Madrid* y *San Luis*) y dos entidades referentes a una materia (*Madre* y *Jaqueca*) que se han enlazado con cuatro conceptos de LEMb. Se puede ver que la precisión no es perfecta ya que solo el 50% de las entidades se emparejan con conceptos correctos de LEMb. Sin embargo, una verificación manual exhaustiva de las 1.075 páginas donde se han extraído un total de 2.658 entidades mejora esta precisión, obteniéndose que el 79,21% las entidades extraídas se emparejaron correctamente con un concepto de LEMb. De todas formas, hay que tener en cuenta que esta precisión es mejorable ya que de momento no se ha incluido ningún mecanismo desambiguación. Simplemente, se ha escogido el primer concepto de LEMb que consigue la mejor valoración en la correspondencia textual de Lucene para las cuatro etiquetas mencionadas en la sección 2.2.

Tipo Entidad	*Texto*	*Concepto LEMb*
Lugar	Madrid	Madrid
	San Luis	San Luis Potosí (Estado)
Materia	Madre	Madres trabajadoras
	Jaqueca	Cefalalgia

Tabla II. Ejemplo de anotación de una página del portal (http://goya.unizar.es/Repositorio/Diplomatario/108.html)

4. Conclusiones

Este trabajo ha presentado un método sencillo para recomendar un modelo KOS entre un conjunto de modelos KOS candidatos, y cómo utilizar el modelo recomendado para la anotación de las páginas de un portal. Para la recomendación del modelo KOS se ha reaprovechado el potencial de modelos clásicos de recuperación de información, véase el modelo vectorial, con el objetivo de ver la correspondencia entre los contenidos que se quieren anotar y el contenido textual de los modelos KOS. Para la parte de anotación de las páginas se ha reutilizado el software disponible de reconocimiento de entidades nombradas y se ha hecho uso del lenguaje SPARQL combinado

con motores de indexación textual como Lucene para encontrar correspondencias entre entidades nombradas y etiquetas textuales asociadas a los conceptos del modelo KOS elegido.

Para verificar la viabilidad de este modelo se ha implementado un prototipo inicial y se ha aplicado sobre el portal InfoGoya. Aunque algunos conceptos extraídos automáticamente del modelo LEMb pueden no ser demasiado precisos, proporcionan una buena base para que los administradores del portal puedan anotar semánticamente el mismo.

Como trabajo futuro proponemos mejorar nuestro proceso de identificación y análisis de modelos KOS con la integración de bases de datos de modelos KOS de relevancia como BARTOC (Verbundzentale des GBV, 2021) para aquellos casos en los que no se cuente con un trabajo previo de análisis de modelos KOS utilizados en un dominio particular. Asimismo, estudiaremos otras técnicas de clasificación para asignar el KOS más cercano al contenido de un portal. También se trabajará en la integración de mecanismos de desambiguación que mejoren el emparejamiento entre entidades reconocidas en el texto y los conceptos del modelo KOS. Por último, se plantea estudiar distintas alternativas para integrar las anotaciones como micro-metadatos RDF embebidos en las páginas HTML del portal semantizado.

Agradecimientos

Este trabajo ha sido financiado parcialmente por el Gobierno de Aragón (proyecto T29_20R).

Referencias

Albertoni, R.; De Martino, M.; Quarati, A. (2016). Integrated quality assessment of linked thesauri for the environment. // International Conference on Electronic Government and the Information Systems Perspective. 221–235.

Apache (2017). The Apache Software Foundation. Apache OpenNLP web site. https://opennlp.apache.org/ (2021-03-24)

Comisión Europea (2021). Portal de Europeana. https://www.europeana.eu/ (2021-03-24)

da Silva Lemos, D.L.; Souza, R.R. (2020). Knowledge Organization Systems for the Representation of Multimedia Resources on the Web: A Comparative Analysis. // Knowledge Organization. 47:4, 300-319.

Daiber, J.; Jakob, M.; Hokamp, C.; Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. // Proceedings of the 9th International Conference on Semantic Systems. 121-124.

Díaz-Corona, D.; Lacasta, J.; Latre, M.Á.; Zarazaga-Soria, F.J.; Nogueras-Iso J. (2019). Profiling of knowledge organisation systems for the annotation of Linked Data cultural resources. // Information Systems. 84, 17-28.

DPLA (2021). Portal de la Digital Public Library of America. https://dp.la/ (2021-03-24)

Khalili, A., Auer, S. (2015). WYSIWYM–Integrated visualization, exploration and authoring of semantically enriched un-structured content. // Semantic Web. 6:3, 259-275.

Lacasta, J.; Falquet, G.; Zarazaga-Soria, F.J.; Nogueras-Iso, J. (2016). An automatic method for reporting the quality of thesauri. // Data & Knowledge Engineering. 104, 1-14.

Lacasta, J.; Nogueras-Iso, J.; Zarazaga-Soria, F. J. (2010). Terminological Ontologies: Design, Management and Practical Applications. Springer.

Miles, A.; Brickley, D. (2005) SKOS Core Guide. http://www.w3.org/TR/swbp-skos-core-guide (2021-03-24).

Rodriguez Rocha, O.; Vagliano, I.; Figueroa, C.; Cairo, F.; Futia, G.; Licciardi, C. A.; Marengo, M.; Morando, F. (2015). Semantic annotation and classification in practice. // IT Professional. 17:2, 33-39.

Savoy J. (2002) Report on CLEF-2001 Experiments: Effective Combined Query-Translation Approach. // Peters, C.; Braschler, M.; Gonzalo, J.; Kluck, M. (eds). Evaluation of Cross-Language Information Retrieval Systems. CLEF 2001. Lecture Notes in Computer Science. 2406, 27-43.

Suominen, O.; Mader, C. (2014). Assessing and improving the quality of SKOS vocabularies. // Journal of Data Semantics. 3:1, 47–73.

Tchechmedjiev, A.; Abdaoui, A.; Emonet, V.; Melzi, S., Jonnagaddala, J.; Jonquet, C. (2018). Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator+. // Bioinformatics. 34:11, 1962-1965.

Thessen, A.E.; Parr, C.S. (2014). Knowledge extraction and semantic annotation of text from the encyclopedia of life. // PloS one. 9:3, e89550.

Universidad de Zaragoza (1996). Exposición Virtual InfoGoya ’96. http://goya.unizar.es/ (2021-03-24)

Verbundzentrale des GBV (2021). Basic Register of Thesauri, Ontologies & Classifications (BARTOC). https://bartoc.org/ (2021-06-08).

Whaley, P.; Edwards, S.W.; Kraft, A.; Nyhan, K.; Shapiro, A.; Watford, S.; Wattam S.; Wolffe, T.; Angrish, M. (2020). Knowledge Organization Systems for Systematic Chemical Assessments. // Environmental Health Perspectives. 128:12, 125001.

Enviado: 2021-03-25. Segunda versión: 2021-06-09.
Aceptado: 2021-06-17.