Propuesta de representación del tesauro Eurovoc en SKOS para su integración en sistemas de información jurídica

A proposal for representing the Eurovoc thesaurus with SKOS for its integration in juridical information systems

Mª Luisa ALVITE DÍEZ (1), Beatriz PÉREZ-LEÓN (2), M. Mercedes MARTÍNEZ-GONZÁLEZ (3), Dámaso-Javier VICENTE BLANCO (4)

(1) Área de Biblioteconomía y Documentación, Facultad de Filosofía y Letras, Universidad de León, Campus de Vegazana s/n, 24071 León, luisa.alvite@unileon.es. (2) Departamento de Informática, Universidad de Valladolid, Edificio T.I.T, Campus Miguel Delibes s/n, 47011 Valladolid, bperezle@ infor.uva.es. (3) mercedes@infor.uva.es. (4) Departamento de Derecho Mercantil, Derecho del Trabajo e Internacional Privado, Facultad de Derecho, Plaza de la Universidad, s/n, 47002, Valladolid, damaso@der.uva.es.

Resumen

Los tesauros se han adaptado progresivamente a la edición en entornos digitales y se enfrentan en la actualidad al reto de demostrar su utilidad e integración en la Web semántica, haciendo uso de estándares como RDF y SKOS. El estudio justifica, en primer lugar, la selección del Tesauro Eurovoc para su integración en un sistema de información legal. Se analizan, igualmente, las dificultades para la representación de Eurovoc en SKOS, así como las herramientas disponibles para la adecuada manipulación de tesauros. Finalmente, se reflexiona sobre el estado de la cuestión y se plantean líneas de trabajo futuras.

Palabras clave: Eurovoc. SKOS. Sistemas de información jurídica. Tesauros. Web Semántica.

Abstract

Thesauri have been progressively adapted to edition in digital environments and are now facing the challenge of demonstrating its utility and integration in the Semantic Web, using standards like RDF and SKOS. The study justifies, first, the selection of the Eurovoc Thesaurus for its integration into a legal information system. It also discusses the difficulties of representing Eurovoc with SKOS, as well as the tools available for the proper handling of thesauri. Finally, reflections on the state of the art in the field and suggestions for future research are provided.

Keywords: Eurovoc. Legal Information Systems. Semantic Web. Thesauri. SKOS.

1. Lenguajes controlados y Web Semántica

La evolución de los lenguajes documentales, en general, y de los tesauros, en particular, se ha visto favorecida por el desarrollo continuado de las Tecnologías de la Información, que han permitido la aparición de aplicaciones informáticas que auxilian en la gestión y mantenimiento automatizado de estos lenguajes, han posibilitado la integración de tesauros en Sistemas de Recuperación de Información y, además, gracias a la Web, la edición y consulta en formato electrónico de un considerable número de tesauros disponibles en la red es ya un hecho.

Ahora bien, estamos de acuerdo con Pastor Sánchez, Martínez Méndez y Rodríguez Muñoz (2009) en que la explotación de los tesauros en la Web es muy limitada, las deficiencias en la recuperación de información en Internet exigen vocabularios interoperables y multilingües. De ahí que estos autores insistan en que el propio concepto de tesauro ha evolucionado para adaptarse a los modelos de representación de la Web semántica, abandonando el paradigma léxico a favor de un paradigma conceptual. Esta adaptación permitiría, según Sánchez-Jiménez y Gil-Urdiciain (2007) superar las deficiencias de las relaciones fuertemente ancladas a nivel léxico que caracterizan a los tesauros tradicionales.

Estas exigencias han conducido a un esfuerzo a nivel internacional por desarrollar normativas que adapten las tradicionales herramientas terminológicas al entorno digital. El trabajo de García Marco et al. (2007) resume este proceso y la labor del Grupo de Vocabularios Controlados para la Recuperación de Información del Comité 50 de la Asociación Española de Normalización (AENOR) para promover en el Marco de AENOR la adecuación de la normativa española sobre tesauros a las nuevas tendencias internacionales, lideradas por la iniciativa americana ANSI/NISO Z39.19-2005 y la propuesta por el

British Standard Institute*, BS 8723: Structured Vocabularies for Information Retrieval*. Se trata, apostilla el estudio, de establecer mecanismos de interconexión entre los distintos lenguajes documentales y conectarlos con los desarrollos que se están llevando a cabo en el ámbito de las ontologías.

A este necesario esfuerzo normalizador se ha unido el impulso de la Web semántica en la que necesariamente han de incorporarse los sistemas de organización del conocimiento como tesauros, vocabularios controlados, topic maps, esquemas de clasificación y otras herramientas conceptuales. Ahora bien, esta Web semántica requiere que los datos sean legibles y comprensibles por los agentes software, para lo cual son necesarios estándares que permitan la representación de información en formatos interoperables.

Con este fin surgen los distintos estándares de representación, comenzando con XML, y siguiendo con los que se apoyan en él. RDF permite representar metadatos. RDF *Schema* añade la posibilidad de representar el vocabulario utilizado en grafos RDF y relacionarlo mediante estructuras clasificatorias sencillas como taxonomías.

OWL (W3C, 2004) aporta el vocabulario estándar para representar ontologías, las herramientas conceptuales más potentes de las que disponemos para expresar semántica al más alto nivel. Las ontologías han adquirido un papel relevante y proliferan en todo tipo de sistemas de información como herramientas que soportan las búsquedas conceptuales y otra serie de funcionalidades. García, Pareja y Pradana (2008) resumen los beneficios adicionales que aporta la formalización del conocimiento como ontología y no como tesauro.

No obstante, sin soslayar su complejidad, no parece tan claro que las ontologías siempre cumplan la función de integración para la que podrían servir, dado que en muchos casos se diseñan *ad-hoc* para un propósito específico y nunca más son reutilizadas. Esto es, su función se limita al sistema local en cuyo contexto se crearon.

Por su parte, SKOS es un "modelo de datos diseñado para compartir sistemas de organización de conocimiento en la Web" (W3C, 2009b). SKOS se puede implementar sobre RDF, utilizarse aisladamente, o combinarlo con OWL. Esta capacidad de interacción con otras propuestas creemos que dota al modelo de una enorme potencialidad.

En la versión básica de SKOS, los recursos conceptuales, esto es, los conceptos, se identifican mediante URI, cada concepto tiene al menos un término que lo representa; los conceptos se relacionan entre si, creando jerarquías de conceptos, y se agregan bajo estructuras denominadas esquemas de conceptos (concept schemes). En su versión avanzada se introduce la posibilidad de agrupar conceptos que están en distintos esquemas de conceptos en colecciones (collections), que pueden estar o no ordenadas.

Por último, e igualmente trascendental, también en el marco de la Web Semántica se proporcionan los lenguajes de consulta estándar que deben permitir recuperar la información modelada utilizando estos estándares de representación. XQuery permite consultar datos XML y SPARQL es un lenguaje de consulta para RDF que se estabilizó como recomendación del W3C a principios de 2008 (W3C, 2008c).

En este marco, parece evidente que la integración en el nivel de conocimiento debe venir de la mano de la utilización combinada de estos instrumentos (herramientas conceptuales y estándares de representación), e incluso que su uso debería ser suficiente para garantizarla. Trataremos de analizar esta premisa en un dominio de aplicación en el que venimos trabajando desde hace tiempo: los sistemas de información jurídica (Martínez González et al., 2009).

2. Eurovoc

Eurovoc es un tesauro multilingüe cuyo origen data de finales de los años setenta, momento en el que el Parlamento Europeo y la Oficina de Publicaciones Oficiales de las Comunidades Europeas deciden trabajar en la creación de un lenguaje documental común (Maciá, 1995).

Es un tesauro multidisciplinar que trata de abarcar todos los campos de actividad de las Comunidades y los propios de un Parlamento. Eurovoc está estructurado en 21 campos temáticos (dominios) y 127 microtesauros (subdominios). Algunos descriptores de los campos temáticos 72 (Geografía) y 76 (Organizaciones internacionales) son polijerárquicos, tienen varios términos genéricos de primer nivel diferentes.

Eurovoc contiene 6645 descriptores, 519 de los cuales son *top terms*, 7756 no descriptores, 6669 relaciones jerárquicas recíprocas y 891 notas. Datos cuantitativos que muestran el volumen y riqueza conceptual de esta herramienta terminológica. Eurovoc respeta las normas ISO 2788-1986 e ISO 5564-1985. Desde enero de 2009 está disponible la versión 4.3 (Eurovoc Thesaurus, 2009).

2.1. Integración del Tesauro Eurovoc en la Herramienta de Información Legal

La línea fundamental de trabajo de este equipo de investigación se ha centrado en el uso de XML como el estándar más adecuado para representar la información jurídica (Martínez, Fuente y Derniame, 2003). Diseñamos esquemas de representación de los textos jurídicos y sus metadatos y utilizamos los estándares asociados a XML para representarlos y acceder a ellos desde diversas aplicaciones (Martínez González et al., 2009).

Una interesante utilidad para los usuarios de estos sistemas es la posibilidad de anotar los textos, o los elementos de información representativos (elementos de estructura en nuestro caso), con comentarios sobre su campo de aplicación, tema o temas con los que está relacionado y otras observaciones. En el caso de herramientas docentes, como algunas de nuestras aplicaciones, esto puede servir para que el profesor indique a los alumnos con qué tema o temas de la materia debería relacionar el documento legal que está analizando. Igualmente se articula la posibilidad más tradicional de clasificar los documentos en función de materia o materias, de modo que se permitan *a posteriori* búsquedas de documentos relacionados con una materia dada. Esta extensión a través de anotaciones es la que motiva nuestra búsqueda de una herramienta donde estén representados los conceptos que utilizarán nuestros usuarios, así como nuestra indagación en los estándares referidos en la sección anterior para representar y manipular convenientemente este conocimiento.

En nuestro caso decidimos utilizar como herramienta conceptual el tesauro Eurovoc mantenido por la Oficina de Publicaciones de las Comunidades Europeas. Se tuvieron en cuenta varias ventajas. En primer lugar, el hecho de que sea el tesauro oficial que la Oficina de Publicaciones utiliza en sus sistemas de información hace de él el más robusto y estándar de los tesauros que se usan en el campo de la información jurídica. Los recursos disponibles para su mantenimiento, entre los cuales es esencial la actualización terminológica que conlleva la inclusión y eliminación de nuevos términos en versiones sucesivas, garantizan que Eurovoc es un tesauro 'vivo', en proceso de adaptación continua a la evolución de los ámbitos de actividad de las instituciones comunitarias y de las distintas lenguas empleadas.

Su utilización por parte de un buen número de organismos oficiales en los distintos países de la Unión Europea, como el Congreso y el Senado en España o numerosos parlamentos autonómicos (Asturias, Castilla y León, Comunidad Valenciana y otros), induce su utilización en cuantos sistemas buscan interoperabilidad con cualquiera de ellos.

Por último, este tesauro se puede usar libremente, en el marco de un convenio con la Oficina de Publicaciones, en el cual se aceptan las normas que subrayan el debido reconocimiento al origen del tesauro, así como las instrucciones que se deben seguir en el caso de que se proponga la extensión del tesauro con nuevos términos (la eliminación de términos está en principio restringida, de tal modo que sólo se decide eliminar términos cuando se publica una nueva versión).

3. Representación de EUROVOC en SKOS

Tras la firma del Convenio entre la Universidad de Valladolid y la Oficina de Publicaciones Oficiales de las Comunidades Europeas nos fue entregada una copia de Eurovoc (versión 4.2) en CD-ROM. El tesauro se organiza en un conjunto de ficheros XML, con sus correspondientes DTDs, sencillos en su estructura interna, aunque complejos en lo que se refiere al número y organización de los ficheros. Se trata, por tanto, de un formato propietario XML-Eurovoc de la Oficina de Publicaciones que no emplea ningún estándar de los mencionados en el Epígrafe 1.

Los pasos siguientes se dirigieron, de un lado, a estudiar posibles APIs (Interfaces de Programación de Aplicaciones) para la gestión de tesauros que nos pudiesen facilitar el desarrollo de aplicaciones y, de otro, a estudiar la viabilidad de usar el formato propietario XML-Eurovoc recibido o a adaptar el mismo a SKOS y proponer una representación en el marco de la Web semántica.

3.1. Análisis de representaciones

En aras de buscar la interoperabilidad, fue descartada la alternativa de almacenar el tesauro en el formato XML-Eurovoc recibido, ya que al tratarse de un formato propietario, su empleo cerraba la posibilidad de realizar importaciones de otros tesauros representados con otros estándares. Así pues, se optó por representar Eurovoc conforme a la propuesta SKOS.

La representación de Eurovoc con SKOS ha sido planteada desde perspectivas diversas. Así, Polo, Álvarez y Rubiera (2008), optan por representar los campos temáticos y microtesauros como esquemas conceptuales (*concept schemes)* y los enlazan mediante propiedades OWL creadas *ad-hoc*. Por su parte, Faro et al.

(2008), representan los campos temáticos como colecciones de esquemas conceptuales (*collections*) y los microtesauros como esquemas conceptuales. Ambas propuestas se ajustan a los *working draft* anteriores al año 2009 (W3C, 2008a y W3C, 2008b).

El análisis de las representaciones SKOS de Eurovoc disponibles evidencia la diversidad de soluciones, que, en nuestra opinión, obedecen a las diferencias entre versiones achacables a una Recomendación del W3C que aún no era estable. Así, existen diferencias significativas entre la versión de SKOS de 2008 y la última Recomendación, de agosto de 2009 (W3C, 2009b). Una de estas discordancias está relacionada con el empleo de las nociones de *Collection* y *Concept scheme* de SKOS, que son utilizados para representar los microtesauros de Eurovoc. La Recomendación de 2009 introduce una variación crucial con respecto a los *Working drafts* previos, no permitiendo que una Colección tenga como elementos a Esquemas conceptuales. En consecuencia, alguna de las propuestas de representación analizadas no se adapta a esta Recomendación en vigor.

3.2. APIs

Existen varias herramientas que permiten manipular tesauros (Ferreyra, Lacasta et al., 2007). Se trata de aplicaciones de usuario que permiten crear y mantener tesauros, pero no aportan una API que se pueda utilizar desde otras aplicaciones.

3.3. Dificultades en la representación

Teniendo presentes las lecciones aprendidas del estudio de las propuestas de representación realizadas sobre los *Working drafts* anteriores a la Recomendación estable de SKOS, hemos decidido realizar una propuesta de representación de Eurovoc atendiendo a la Recomendación de agosto de 2009.

Dos han sido los principales problemas hallados en la representación del tesauro Eurovoc que dificultan la generación automática sencilla bajo SKOS. De un lado, la coincidencia de códigos identificadores entre descriptores, campos temáticos y microtesauros que impide usar dichos códigos directamente para la obtención de URIs únicas. De otro lado, la polijerarquía en Eurovoc a la que hemos aludido en el punto anterior.

Se ha optado, para la gestión de las URIs, por efectuar una modificación en las reglas de generación de Eurovoc SKOS, de modo que, cuando se trate de un descriptor, la URI generada incluya un identificador alfabético que indique que se trata inequívocamente de un descriptor. En el caso de los descriptores polijerárquicos de los dominios 72 y 76 la aplicación mostrará todos los términos genéricos de primer nivel que le correspondan al descriptor.

3.4. Resultados

El análisis de las propuestas de representación analizadas nos lleva a plantearnos las siguientes cuestiones: ¿Son compatibles entre ellas estas soluciones? ¿Serán no obstante compatibles entre ellas las nuevas propuestas que se ajusten a la nueva Recomendación? Este es un aspecto relevante si tenemos en cuenta que los tesauros pueden evolucionar con el tiempo, o si extendemos la experiencia a otros tesauros que no estén tan estrictamente controlados en su mantenimiento por algún organismo, como es el caso de Eurovoc, y pudieran ser objeto de modificaciones con más facilidad. La incompatibilidad entre la representación de una versión y la posterior nos situaría una vez más ante un problema de interoperabilidad.

Asimismo, la elección entre *Collection* y *Concept scheme*, dos estructuras diferentes que SKOS ofrece para representar dominios y subdominios de los tesauros, introduce heterogeneidad en un nivel superior, que no pueden resolver los lenguajes de consulta apropiados, como SPARQL.

4. Conclusiones y líneas de trabajo

La integración de las herramientas de representación de conocimiento utilizadas en distintos sistemas de información es un reto vinculado estrechamente a la Web Semántica.

No obstante, la variedad de propuestas de representación encontradas para el tesauro Eurovoc, y los problemas comentados en los puntos anteriores, demuestran que la integración de conocimiento, tanto a nivel conceptual como formal, encuentra más obstáculos de lo que en principio podría parecer, incluso cuando se utilizan estándares para la representación de semántica.

En este trabajo se ha presentado un caso de aplicación, en el que se pretende utilizar un tesauro 'estándar', y se han analizado los posibles problemas de integración que pueden surgir, tanto en el nivel conceptual como en el de representación de los tesauros. Tratamos de concluir en el momento actual la representación de Eurovoc en SKOS atendiendo a la Recomendación de agosto de 2009 (W3C, 2009b), conscientes de que se requeriría una mayor

claridad sobre cómo modelar dominios y subdominios en el estándar.

En nuestra opinión una API pública de uso general para tesauros que utilice los estándares de representación del W3C simplificaría el desarrollo de sistemas que utilizan estas herramientas conceptuales.

Por último, esperamos completar el desarrollo de una API genérica y del software que la implemente como biblioteca para poder realizar pruebas sobre la aplicación docente empleada en la asignatura de Derecho Internacional Privado (Universidad de Valladolid), que nos permitiría, tras la etapa de revisión y mejoras, su presentación pública.

5. Referencias

Eurovoc Thesaurus (2009). Luxemburgo: Oficina de Publicaciones Oficiales de las Comunidades Europeas, 2009. http://europa.eu/eurovoc (2010-04-10)

Faro, Sebastiano; Francesconi, Enrico; Marinai, Elisabetta; Sandrucci, V. (2008). EUROVOC Studies LOT2 D2.3. Report on execution and results of the interoperability tests. // Tech. Rep. 10118. Florencia: Publications Office of the EC, Institute of Legal Information Theory and Techniques ITTIG (January 2008).

Ferreyra, D. TemaTres: software libre para gestión de tesauros. http://www.r020.com.ar/tematres/index.html (2010-03-04).

García Marco, Francisco Javier (Coord.); Agustín Lacruz, María del Carmen; Caro Castro, Carmen; Martínez Usero, José Ángel; San Segundo Manuel, Rosa (2007). Proyectos internacionales de reforma y ampliación de las normas sobre tesauros para su adaptación a los nuevos contextos de integración e interoperabilidad en el entorno digital. // VIII Congreso ISKO – España. León: Universidad de León, 2007. 389-398.

García Torres, Alberto; Pareja Lora, Antonio; Pradana López, Daniel (2008). Reutilización de tesauros: el documentalista frente al reto de la Web semántica. // El Profesional de la Información. 17:1 (Enero-Febrero 2008) 8-21.

Lacasta, J.; Nogueras, J.; López-Pellicer, F. J.; Muro-Medrano, p. ; Zarazaga-Soria, F. (2007). ThManager: An Open Source Tool for creating and visualizing SKOS. // Information Technology and Libraries (ITAL). 26:3 (2007) 39-51.

Maciá, Mateo (1995). El tesauro EUROVOC. // Revista General de Información y Documentación. 5:2 (1995) 265-284. http://revistas.ucm.es/byd/11321873/articulos/RGID9595220265A.PDF. (2010-03-21).

Martínez González, M. M.; Vicente Blanco, D.J.; Fuente Redondo, p. de la; Adiego Rodríguez, J.; Pisabarro Marrón, A. M.; Sánchez Felipe, J. M. (2009). Estructura, semántica, extracción de información y XML legislativo: experiencias en la Universidad de Valladolid. // Scire: Representación y Organización del Conocimiento. 15:1 (Enero-Junio 2009) 173-186.

Martínez, M. M.; Fuente, p. de la; Derniame, J.C. (2003). XML as a means to support information extraction from legal documents. // International Journal of Computer Systems Science and Engineering. 18:5 (September 2003) 263-277.

Pastor-Sánchez, Juan-Antonio; Martínez Méndez, Francisco Javier; Rodríguez-Muñoz, José Vicente (2009). Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. // Information Research. 14:4 (December 2009). http://informationr.net/ir/14-4/paper422.html (2010-03-15).

Polo Paredes, Luis; Álvarez Rodríguez, José María; Rubiera Azcona, Emilio (2008). Promoting Government Controlled Vocabularios for the Semantic Web: the EUROVOC Thesaurus and the CPV Product Classification System. // Semantic Interoperability in the European Digital Library: Proceedings of the First International Workshop, SIEDL 2008, Tenerife, June 2, 2008. Lecture Notes in Computer Science. 50212 (2008) 111-122.

Sánchez-Jiménez, Rodrigo; Gil Urdiciain, Blanca (2007). Lenguajes documentales y ontologías. // El Profesional de la Información. 16:6 (Noviembre-Diciembre 2007) 551-560.

World Wide Web Consortium (W3C) (2004). OWL Web Onthology Language guide. W3C Recommendation 10 Feb. 2004, 2004. http://www.w3.org/TR/owl-guide (2010-04-10).

World Wide Web Consortium (W3C). (2008a). SKOS Simple Knowledge Organization System Primer. W3C Working Draft 29 August 2008, 2008. http://www.w3.org/TR/2008/WD-skos-primer-20080829/ (2010-04-10).

World Wide Web Consortium (W3C). (2008b). SKOS Simple Knowledge Organization System Reference. W3C Working Draft 29 August 2008, 2008. http://www.w3.org/TR/2008/WD-skos-reference-20080829/. (2010-04-10).

World Wide Web Consortium (W3C). (2008c). SPARQL Query Language for RDF. W3C Recommendation 15 January 2008, 2008. http://www.w3.org/TR/owl-guide (2010-04-10).

World Wide Web Consortium (W3C). (2009a). SKOS Simple Knowledge Organization System Reference. W3C Proposed Recommendation 15 June 2009, 2009. http://www.w3.org/TR/2009/PR-skos-reference-20090615/ (2010-04-10).

World Wide Web Consortium (W3C). (2009b). SKOS Simple Knowledge Organization System Reference. W3C Recommendation 18 August 2009, 2009. http://www.w3.org/TR/2009/REC-skos-reference-20090818/ (2010-04-10).

Recibido: 2010-04-13. Aceptado: 2010-05.24.