# SKOS en la integración de conocimiento en los sistemas de información jurídica

M. Mercedes Martínez-González[1], Beatriz Pérez-León[1], and M. Luisa Alvite-Díez[2]

[1] Departamento de Informática, Universidad de Valladolid
Edificio T.I.T., Campus 'Miguel Delibes' s/n, 47011 Valladolid
{mercedes,bperezle}@infor.uva.es
[2] Área de Biblioteconomía y Documentación, Universidad de León
Facultad de Filosofía, Campus de Vegazana, s/n, 24071 León
luisa.alvite@unileon.es

**Resumen** Los tesauros son herramientas conceptuales que representan un área de conocimiento. Los estándares RDF y SKOS, vinculados a la Web Semántica, permiten su representación con lenguajes y vocabularios estándares, lo cual facilita la integración de conceptos. A pesar de la importante ventaja que supone disponer de estos estándares, no es suficiente para garantizar totalmente la integración de las herramientas conceptuales utilizadas. Demostramos esta afirmación en el caso de los sistemas de información jurídica, en los cuales analizamos el uso del tesauro Eurovoc y su representación con SKOS: propuestas, existencia de herramientas estándares para tesauros, y dificultades para la integración. A partir de este análisis se plantean reflexiones sobre el estado de la cuestión en el momento actual.

**Palabras clave:** SKOS, tesauros, Eurovoc, sistemas de información jurídica, Web Semántica

## 1. Estándares para la integración de conocimiento

Una de las bases de la integración en los sistemas de información en el nivel de conocimiento ha sido y es la utilización de ontologías y otras herramientas conceptuales similares, como vocabularios controlados, tesauros o topic maps.

La incorporación de los sistemas de organización del conocimiento (*Knowledge Organization Systems, KOS*) como tesauros, vocabularios controlados o esquemas de clasificación a los nuevos entornos tecnológicos ha experimentado un gran empuje gracias a la Web Semántica. Las ontologías son las herramientas conceptuales más potentes de las que disponemos para expresar semántica al más alto nivel. Proliferan en todo tipo de sistemas de información como herramientas que soportan las búsquedas conceptuales y otra serie de funcionalidades. No obstante, no está tan claro que siempre cumplan la función de integración para la que podrían servir, ya 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.

La utilización compartida de un conjunto de conceptos con semántica bien definida (conceptos que se representan mediante un conjunto de términos representantes) sería la situación ideal –reutilización de las mismas herramientas conceptuales–. Pero, como hemos dicho, no es la situación habitual, ya que en muchos casos cada sistema utiliza sus propios conceptos. La integración semántica es la encargada de paliar este problema.

La Web Semántica no sería tal si no se cumple el objetivo de que los datos sean legibles y comprensibles por los agentes software [2], 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 aporta el vocabulario estándar para representar ontologías. SKOS es un "modelo de datos diseñado para compartir sistemas de organización de conocimiento en la Web"[17]. SKOS se puede implementar sobre RDF, utilizarse aisladamente, o combinarlo con OWL [17].

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.

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 [4] permite consultar datos XML, y SPARQL [15] es un lenguaje de consulta para RDF que se estabilizó como recomendación del W3C a principios de 2008.

Parece claro en este contexto 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. Sin embargo, esta última afirmación debería ser más bien una pregunta. Una pregunta cuya respuesta vamos a analizar en un dominio de aplicación con el que venimos trabajando desde hace tiempo [9]: los sistemas de información jurídica.

## 2. Representación de conocimiento en los sistemas de información jurídica mediante tesauros y XML

En la manipulación de información jurídica el uso de XML como el estándar más adecuado para representar la información se ha generalizado de tal modo que es imposible hablar de sistemas de información jurídica sin pensar en la utilización de XML [1,3]. Nuestro trabajo viene siguiendo esta filosofía desde hace tiempo [10]. 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 [9].

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 [13], esto puede servir para que el profesor indique a los alumnos con qué tema o temas de la materia debería relacionar el documento que está analizando. También está, por supuesto, la posibilidad más tradicional de clasificar los documentos en función de materia o materias, de modo que después se permitan búsquedas de documentos relacionados con una determinada materia. 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 vistos en la sección 1 para representar y manipular convenientemente este conocimiento.

La primera cuestión que nos planteamos fue: ¿existe una herramienta de representación de conocimiento que podamos reutilizar o necesitaremos crear nuestra propia herramienta conceptual? En nuestro caso decidimos utilizar el tesauro Eurovoc [12], mantenido por la Oficina de Publicaciones de las Comunidades Europeas[3]. Las ventajas de este tesauro son varias. 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', que no queda obsoleto con el paso del tiempo. Su utilización por parte de un buen número de instituciones oficiales en los distintos países de la Unión Europea, como el Senado en España [5], 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).

Eurovoc es un tesauro multilingüe cuyo origen se remota a 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. Eurovoc está estructurado en 21 campos temáticos (dominios) y 127 microtesauros (subdominios). Contiene 6045 descriptores, 7756 no descriptores, 6669 relaciones jerárquicas recíprocas y 891 notas, datos cuantitativos que muestran el volumen y valor conceptual de esta herramien-

[3] En adelante *Oficina de Publicaciones*

ta terminológica. Eurovoc respeta las normas ISO 2788-1986 e ISO 5564-1985. Desde enero de 2009 está disponible la versión 4.3.

Tras firmar el correspondiente convenio entre la Universidad de Valladolid y la Oficina de Publicaciones, recibimos una copia del tesauro en un CD-ROM. El tesauro se distribuye en un conjunto de ficheros XML, con sus correspondientes DTD, sencillos en su estructura interna, aunque complejos en el número y organización de los ficheros. No se utiliza SKOS ni ningún otro estándar de representación de los vistos en la sección 1, razón por la cual nos vamos a referir en adelante a este formato como el 'formato propietario XML-Eurovoc de la Oficina de Publicaciones'.

Una vez disponíamos de una copia del tesauro, los pasos siguientes son buscar una API para tesauros que nos facilite el desarrollo de aplicaciones, y decidir si vamos a utilizar Eurovoc en el formato en que lo hemos recibido o preferimos una adaptación a SKOS, que parece especialmente indicado para representar tesauros en el marco de la Web Semántica [16].

## 3. Manipulación y representación del tesauro Eurovoc

Existen varias herramientas que permiten manipular tesauros [7,8]. No obstante, hay varios inconvenientes para el uso que pretendemos. Buscamos una herramienta que nos permita abstraernos de las particularidades del almacenamiento del tesauro y cómo está representado. Sin embargo, las herramientas encontradas son aplicaciones de usuario, que permiten crear o manipular tesauros, pero no aportan una API que se pueda utilizar desde otras aplicaciones, que es lo que buscamos. Por otro lado, no es posible importar directamente el tesauro Eurovoc tal cual lo hemos recibido. Esto nos sitúa ante la necesidad de ocuparnos del almacenamiento del tesauro y su gestión desde nuestras aplicaciones. La solución por la que optamos es desarrollar una API genérica y utilizarla desde nuestras aplicaciones.

Almacenar el tesauro supone escoger un formato de representación, para lo cual consideramos varias alternativas. La primera es almacenar el tesauro en el formato XML en que lo hemos recibido. Esta opción se descartó, ya que al ser un formato propietario, su utilización cierra la posibilidad de importar directamente cualquier otro tesauro que no esté representado con este mismo formato, lo cual incluye tesauros representados con estándares como SKOS. Así pues, para facilitar la interoperabilidad, optamos por utilizar una representación estándar para los tesauros, conforme con la propuesta SKOS.

La representación de Eurovoc con SKOS ha sido abordada en varias ocasiones, desde distintos puntos de vista, representando los dominios y microtesauros como esquemas conceptuales (*concept scheme*) y enlazándolos mediante propiedades OWL creadas ad-hoc [11] o representando los dominios como colecciones de esquemas conceptuales, que serían los microtesauros [6]. De ambas propuestas, que se ajustan a los *working draft* anteriores al año 2009 [14], hemos estudiado su posible reutilización.

A pesar de que todas ellas se ajustan a SKOS, son diferentes. Una de las razones de esta diversidad en las soluciones es el hecho de que SKOS no es aún una Recomendación estable. De hecho, existen varias diferencias entre la versión de SKOS de 2008 y la última Recomendación propuesta, de junio de 2009 [17]. Una de ellas está relacionada precisamente con la utilización de los conceptos *Collection* y *ConceptScheme* de SKOS, que son utilizados para representar los *microtesauros* de Eurovoc. La *Proposed Recommendation* de junio de 2009 introduce una variación respecto a los *working draft* previos, de modo que ya no se permite que una colección tenga como elementos a esquemas conceptuales. Esto implica que algunas de las propuestas analizadas quedarán obsoletas una vez se consolide la nueva versión como Recomendación estable.

Por otro lado, el análisis de estas propuestas también nos lleva a plantearnos las siguientes cuestiones: ¿Son compatibles entre ellas estas propuestas? ¿Serán no obstante compatibles entre ellas las nuevas propuestas que se ajusten a la nueva recomendación? Este es un aspecto importante si tenemos en cuenta que los tesauros pueden evolucionar con el tiempo, o si extendemos la experiencia a otros tesauros que no estén tan férreamente controlados en su evolución por algún organismo como es el caso de Eurovoc, en cuyo caso pueden sufrir modificaciones aún con más facilidad. La incompatibilidad entre la representación de una versión y la posterior nos situaría una vez más ante un problema de interoperabilidad.

Finalmente, si pensamos en una API genérica para tesauros, la cuestión que surge es si la elección de una u otra representación condiciona la generalidad de una API y sus implementaciones. La heterogeneidad sintáctica puede solventarse con los lenguajes de consulta apropiados, como SPARQL. Pero la elección entre *Collection* y *ConceptScheme*, 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.

## 4. Conclusiones

La integración de las herramientas de representación de conocimiento utilizadas en distintos sistemas de información es un reto directamente relacionado con la Web 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.

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. No obstante, la variedad de propuestas de representación encontradas para el tesauro Eurovoc, y los problemas comentados en las secciones anteriores, demuestra 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.

## Referencias


1. Alvite Díez, M. L. Las bases de datos jurídicas y el uso del lenguaje XML en España. *SCIRE. Representación y Organización del Conocimiento 15*, 1 (Enero-Junio 2009), 33–57. ISSN 1135-3716.
2. Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. *Scientific American* (May 2001).
3. Biagioli, C., Francesconi, E., and Sartor, G., Eds. *Proceedings of the V Legislative XML Workshop, Florencia, Italia, 14-16 June 2006* (2007), European Press Academic Publishing.
4. Boag, S., Chamberlin, D., Fernández, M. F., Florescu, D., Robie, J., and Siméon, J. XQuery 1.0: An XML Query Language. W3C Recommendation 23 January 2007, W3C, the World Wide Web Consortium, http://www.w3.org/TR/2007/REC-xquery-20070123/, Jan. 2007.
5. Cueto Aparicio, M. Eurovoc thesaurus use at the Senate of Spain. In *Information in Africa project workshop* (27-30 June 2006).
6. Faro, S., Francesconi, E., Marinai, E., and Sandrucci, V. EUROVOC Studies LOT2 D2.3 -Report on execution and results of the interoperability tests. Tech. Rep. 10118, Publications Office of the EC, Institute of Legal Information Theory and Techniques ITTIG, Jan. 2008.
7. Ferreyra, D. TemaTres: software libre para gestión de tesauros. URI: http://www.r020.com.ar/tematres/index.html.
8. Lacasta, J., Nogueras, J., López-Pellicer, F. J., Muro-Medrano, P., and Zarazaga-Soria, F. ThManager: An Open Source Tool for creating and visualizing SKOS. *Information Technology and Libraries (ITAL) 26*, 3 (2007), 39–51.
9. Martínez González, M. M., Vicente Blanco, D.-J., de la Fuente Redondo, P., Adiego Rodríguez, J., Pisabarro Marrón, A. M., and Sánchez Felipe, J. M. 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. ISSN 1135-3716.
10. Martínez, M. M., de la Fuente, P., and Derniame, J.-C. XML as a means to support information extraction from legal documents. *International Journal of Computer Systems Science and Engineering 18*, 5 (Sept. 2003), 263–277.
11. Paredes, L. P., Maria, J., Rodriguez, A., and Azcona, E. R. Promoting Government Controlled Vocabularios for the Semantic Web: the EUROVOC Thesaurus and the CPV Product Classification System. In *The Fifth European Semantic Web Conference, 1-5 june 2008, Tenerife, Spain* (2008), vol. 50212 of *Lecture Notes in Computer Science*, Springer, pp. 111–122.
12. Publications Office. *Eurovoc thesaurus*. Accesible en http://europa.eu/eurovoc (última consulta: 10/07/2009).
13. Vicente, D.-J., Martínez, M., Sánchez, J. M., and Adiego, J. Experiences on teaching international private law with the support of an e-learning tool. In *The LEFIS virtual campus design* (Albarracín (España), May 2007). Avaible online at http://www.lefis.org/meetings/workshops/2007/albarracin_2007/contenido/albarracin2007_damaso.ppt.
14. W3C, the World Wide Web Consortium. *SKOS Simple Knowledge Organization System Reference. W3C Working Draft 29 August 2008*, Aug. 2008. http://www.w3.org/TR/2008/WD-skos-reference-20080829/.

15. W3C, the World Wide Web Consortium. *SPARQL Query Language for RDF. W3C Recommendation 15 January 2008*, Jan. 2008. http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.
16. W3C, the World Wide Web Consortium. *SKOS Simple Knowledge Organization System Primer*, June 2009. W3C Working Draft. http://www.w3.org/TR/2009/WD-skos-primer-20090615/.
17. W3C, the World Wide Web Consortium. *SKOS Simple Knowledge Organization System Reference. W3C Proposed Recommendation 15 June 2009*, June 2009. http://www.w3.org/TR/2009/PR-skos-reference-20090615/.