

Definición de una metodología para la construcción de Sistemas de Organización del Conocimiento a partir de un corpus documental en Lenguaje Natural

Sonia Sánchez-Cuadrado

Universidad Carlos III de Madrid
Avda. Universidad 30, 28911 Leganés
ssanche@ie.inf.uc3m.es

José Antonio Moreira González

Universidad Carlos III de Madrid
C/ Madrid 126, 28903 Getafe
jamore@bib.uc3m.es

Jorge Morato Lara

Universidad Carlos III de Madrid
Avda. Universidad 30, 28911 Leganés
jorge@ie.inf.uc3m.es

Mónica Marrero Linares

Universidad Carlos III de Madrid
Avda. Universidad 30, 28911 Leganés
mmarrero@inf.uc3m.es

Resumen: Se propone una metodología para la construcción automatizada de KOS adaptable a diferentes entornos a partir de un corpus documental y unas aplicaciones de tratamiento textual que soporten todo el proceso de construcción y mantenimiento automatizado del KOS. Esta metodología se ha aplicado a diferentes entornos reales, comprobando que se trata de una metodología adaptable y obteniendo una reducción significativa de la intervención de expertos del dominio.

Palabras clave: metodología, Sistemas de Organización del Conocimiento, KOS, adquisición de conocimiento, sistema PLN, relaciones semánticas.

Abstract: A methodology to automatic KOS construction is proposed based on information extraction from natural language documents. Also, a set of NLP tools have been implemented to help in the development and management process. The methodology has been tested in real world projects. Results show that the methodology is highly adaptable and have a low dependence of domain experts.

Keywords: Methodology, Knowledge Organization Systems, KOS, Knowledge acquisition, NLP tools, semantic relationships.

1 Introducción

El objetivo de esta investigación es proponer una metodología adaptable para la construcción automatizada de Sistemas de Organización del Conocimiento a partir de documentos en lenguaje natural de dominios específicos procedentes de entornos y necesidades reales. Este propósito parte de la premisa de que la mayor parte del conocimiento está explicitado en los documentos de un dominio mediante términos y relaciones y que sólo el conocimiento que no esté expresado en los documentos tendrá que ser aportado por los expertos del dominio.

Para diferentes autores como Hodge (2000) o Zeng y Chan (2004) el término Sistemas de Organización de Conocimiento, también conocido como KOS, engloba diferentes tipos de esquemas para organizar la información y promover la gestión del conocimiento, como esquemas de clasificación y categorización, encabezamientos de materias, archivos de autoridades, tesauros, redes semánticas y ontologías. Actualmente, los KOS representan un área de creciente interés por la variedad de disciplinas que han confluído en la necesidad de disponer de estos recursos. Cada una de las áreas de conocimiento ha propuesto unos sistemas de acuerdo a sus necesidades y que por tanto varían en su denominación y en algunas características aunque subyace un modelo

común (Daconta et al., 2003: 157; Lassila y McGuinness, 2001; Gruninger y Uschold, 2002). Algunas de estas características entre los distintos tipos de KOS son:

- Representación simplificada de la realidad
- Conceptos y relaciones de un dominio
- Estructuras flexibles en riqueza semántica
- Proporcionar un vocabulario normalizado y consensuado

Los KOS suponen un recurso que beneficia la comunicación entre expertos y que permite compartir conocimiento de un dominio o una lengua (ISO 2788:1986; NISO Z39.19: 2005). Además aplicado a la RI se mejora en la clasificación y descripción de documentos mediante términos no ambiguos, y la posibilidad de proporcionar un sistema de expansión y restricción de consultas (Foskett, 1971; Baeza-Yates y Ribeiro-Neto, 1992; Ingwersen, 1992). También se ha aplicado en la Terminología (Cabré, 1993), la Ingeniería del Software mediante el análisis de dominios para la reutilización del software (Prieto-Díaz, 1991; Lloréns, 1996); la Ingeniería Artificial incorporando ontologías que permitan realizar inferencias (Gómez-Pérez, 2003: 119-132); en la Web Semántica mediante la construcción de vocabularios de metadatos (Berners-Lee et al., 2001; Daconta et al, 2003), o incluso como mapas conceptuales para recursos educativos (Novak, 1994; 1998).

Las distintas metodologías relacionadas con la construcción de KOS (Gómez-Pérez et al., 2003) coinciden en que deben cumplir las siguientes características: claridad, coherencia, especificación independiente, extensibilidad, vocabulario mínimo con definiciones y denominaciones normalizadas. Así mismo, a partir de las propuestas, se han detectado unas fases comunes para su construcción como:

- Determinar un ámbito o dominio
- Adquisición del conocimiento
- Comprobación de posibles anomalías e inconsistencias
- Evaluación
- Aplicación
- Mantenimiento

Para algunas de las fases existen iniciativas que utilizan herramientas que contribuyen a realizar estas tareas, no obstante la mayor carga de trabajo recae sobre el experto encargado de la construcción del KOS.

Las propuestas para la construcción manual de KOS (Aitchison et al., 1972: 141; Lancaster,

1986; Van Slype, 1991; Noy y McGuinness, 2001) presentan problemas significativos. Por una parte, los KOS consumen grandes recursos económicos y humanos durante un largo periodo, y además implican un coste extra cada vez que deben ser actualizados. A esto se debe sumar la dificultad para consensuar los diferentes criterios de los expertos para la organización del conocimiento. Pero sin duda, uno de los problemas más preocupantes es la falta de disponibilidad de expertos del dominio y la desmotivación de estos expertos en las fases de construcción y actualización. Por este motivo, los principales puntos débiles se encuentran relacionados con la intervención de los expertos y con la adquisición del conocimiento (Antonioniou y Harmelen, 2004: 211; Gómez-Pérez et al., 2004:107).

Por otro lado, la construcción automatizada de KOS presenta las siguientes dificultades:

1. Definir el tipo de KOS y la estructura de conocimiento. Es frecuente que clientes y usuarios no sepan explicar que características y funcionalidad esperan del KOS.
2. Definir y recopilar el material el conocimiento que se representará en el KOS condiciona directamente los resultados, la dificultad en la construcción de la estructura de conocimiento y la calidad del resultado:
 - Los documentos están en un idioma diferente al que se está procesando
 - Los documentos están en varios idiomas
 - Los documentos son multidisciplinarios
 - Los documentos presentan diferentes grados de especificidad
 - Los documentos no están correctamente escritos (estilo-ortografía)
 - Los documentos presentan sintaxis no formalmente estructurada
 - Problemas para extraer texto de algunos formatos (ej. Texto de imágenes)
3. Definir la funcionalidad de las herramientas informáticas para las fases que pueden ser automatizadas. Existen dos funciones fundamentales: extracción de conocimiento e identificación del conocimiento. La primera debe seleccionar aquella información que pueda aportar conocimiento significativo para una estructura organizativa (por lo tanto una indexación selectiva). Por otra parte, el proceso de indexación tenderá a registrar la

mayor cantidad de información (por lo tanto, una indización exhaustiva).

4. Análisis del resultado del KOS. Se requiere un análisis de los resultados de la estructura de conocimiento construida, debido a que los sistemas de adquisición de información tienden a ser genéricos.

2 Definición de la Metodología

En primer lugar, se establece una definición de roles para la construcción del KOS (Fraga et al, 2006): ingeniero de dominio (ID), experto de dominio (ED) y responsable de dominio (RD) y después una definición de una metodología. Esta metodología estará compuesta por actividades de la construcción del KOS y actividades de apoyo relacionadas con aspectos informáticos, documentación y con el personal experto. La metodología desarrollada (Sánchez-Cuadrado, 2007) utiliza aplicaciones software como ayuda a las distintas fases, pero también como soporte del KOS

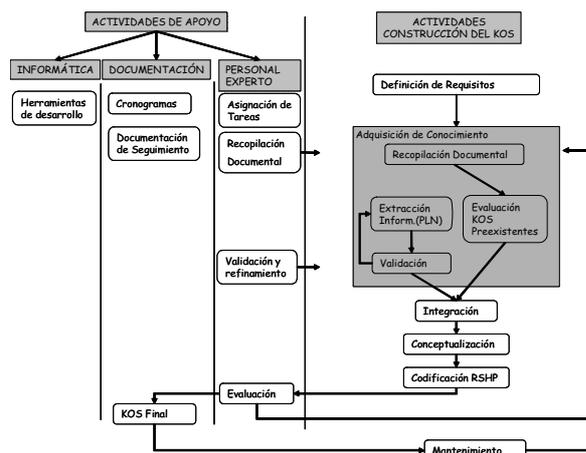


Figura 1: Metodología CAKE para construcción de KOS

Los fundamentos de la metodología CAKE (Figura1) se basan en una serie de actividades para la construcción del KOS (Sánchez-Cuadrado et al., 2006):

1. Definición de requisitos para la identificación del dominio
2. Adquisición de conocimiento
 - Recogida de documentación y filtrado: selección del corpus especializado
 - Propuesta de un conjunto reducido de categorías que sirvan de semilla a la incorporación de otros nodos de la taxonomía inicial

- Identificación del vocabulario de especialidad: extracción, valoración y validación de vocabulario
- Identificación de relaciones de especialidad: extracción, valoración y validación de relaciones

3. Evaluación de la calidad del KOS
4. Mantenimiento del KOS

Las fases de definición de la estructura de conocimiento mediante la definición de requisitos y la definición del corpus documental se realiza mediante: entrevistas con los expertos y la selección de documentos.

1. Entrevistas con los expertos.
 - Determinar el dominio
 - Determinar las preguntas que deberían hacerse a un experto: finalidad, tema, subtemas, preguntas a realizar al sistema RI
 - Dar pautas a los expertos para la construcción del corpus
 - El resultado de esta fase debe ser:
 - Una estructura taxonómica que represente a muy alto nivel los componentes básicos que se desean representar
 - Un listado de preguntas y respuestas que desean resolver para una consulta
 - Un informe de directrices y recomendaciones para la construcción de un corpus
2. Selección de documentos: diferenciar los documentos que están orientados a la construcción de la estructura de conocimiento con los que están orientados a ser documentos de indización.
 - Para la construcción KOS. Es un requisito que estos documentos contengan (aunque sea parcialmente) los términos utilizados en los documentos (cuanto más estructurados los documentos, mejor)
 - Listados de términos que utilicen o de índices de libros o informes que tengan.
 - Si tienen tesauros parciales
 - Glosarios que utilicen (o material de formación de personal)
 - De la entrevista y de los documentos estructurados debería salir un primer esbozo de estructura de conocimiento. Esta debería ser evaluada por un/unos experto/s y confirmar la orientación correcta para que pueda ser ampliada.

Las fases de identificación del vocabulario (3) e identificación de relaciones (4) están basadas en sistemas de PLN (Figura 2) que identifican conceptos (simples y complejos) y relaciones léxico-semánticas a partir de patrones y relaciones sintagmáticas (Sánchez-Cuadrado et al., 2003).

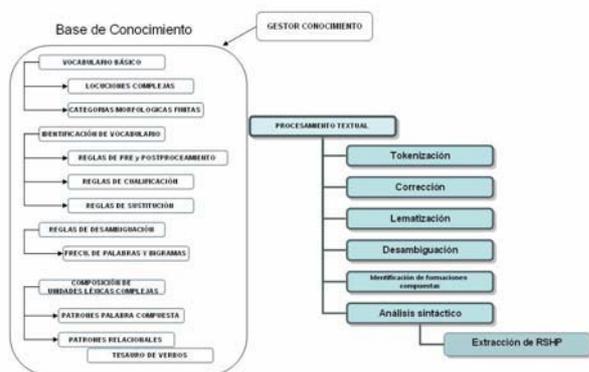


Figura 2: Base de datos de conocimiento y tecnologías lingüísticas aplicadas a la adquisición de conocimiento

Las fases de valoración y validación de los términos y relaciones de especialidad se realizan con una herramienta para la toma de decisiones sobre posibles términos o relaciones conflictivos. Las herramientas que se deben utilizar estarán en función de: la finalidad del sistema, las características del corpus, el volumen del corpus, la implicación de los expertos en el proceso, las técnicas de evaluación y mantenimiento. Por lo tanto se analizará:

- Finalidad del sistema,
- Las características del corpus,
- Procesamiento textual-calidad de los resultados
- El volumen del corpus,
- La implicación de los expertos en el proceso
- Las técnicas de evaluación
- El mantenimiento del KOS

En aquellas tareas que deban ser realizadas por expertos del dominio, las herramientas son sencillas, y el tiempo que los expertos deben dedicar a estas tareas debe ser mínimo. Para lograr esto, la solución pasa por obtener buenos resultados y procesos de filtrado automatizados. En general, el análisis y la valoración de los resultados en las diferentes fases de construcción de estructuras de conocimiento, es

por parte de los expertos del dominio. Por tanto, la presentación de los resultados debe ser clara y lo más concreta posible. Una forma de lograr claridad y concreción será mediante conocimiento contextualizado.

Los procesos de mantenimiento deberán ser coherentes (no repetir información, no insertar información contradictoria, no información errónea, etc.). Uso fácil, y actualización en cascada y coherente.

3 Aplicación de la metodología a entornos reales

Esta propuesta es resultado de la construcción de distintos KOS para entornos reales según la definición de requisitos expresados por la institución.

Esta metodología se ha empleado en el entorno petroquímico (REPSOL-YPF) siendo construidos por separado diferentes áreas de conocimiento de la organización. Se construyeron cuatro KOS aplicando la herramienta de PLN para la automatización de la fase de adquisición del conocimiento y herramientas Web para la toma de decisiones de las fases de valoración y validación de términos y relaciones por miembros de la organización. Los KOS obtenidos para el entorno petroquímico tenían la función de indizar de forma automática para poder recuperar los documentos.

A continuación se muestran algunos de los resultados obtenidos en la aplicación de los métodos propuestos a los diferentes dominios: REPSOL-YPF, SAGE-SP, Oficina Defensor del Pueblo, prototipo de la Guardia Civil en cuanto a metodología automatizada, también se ha aplicado a la creación manual de dominios en el proyecto del Archivo General de la Nación de la República Dominicana (AGN). En todos estos proyectos se ha utilizado el modelo RSHP.

	REPSOL-YPF	SAGE	P-GC	Defensor del Pueblo	AGN
Modelo RSHP	si	si	si	si	si
Categorías Generales	no	si	si	Si	si
Recursos documentales	si	no	no	No	si
Análisis recursos estructurados	no	no	no	Si	si
Análisis recursos semi-estructurados	si	no	no	No	no
Análisis recursos no estructurados	si	si	si	No	no
Extracción de Entidades	no	si	si	No	no

Valoración de términos por la organización	si	no	no	Si	si
Validación de términos por la organización	si	no	no	Si	si
Valoración de relaciones por la organización	si	no	no	Si	si
Validación de relaciones por la organización	si	no	no	Si	si

Tabla 1: Fases de construcción de KOS aplicadas a diferentes entornos

En la presente tabla se presentan estos resultados de forma resumida según se haya realizado o no determinada etapa de la metodología en cada uno de los proyectos (Tabla 1).

En cuanto al establecimiento a priori de una clasificación general, se aplicó en los dominios de SAGE, en el prototipo de la Guardia Civil y en el dominio del AGN, confirmando que facilita no sólo las primeras fases de distribución de los términos en categorías y la facilidad para entender la formación del dominio, si no la definición de relaciones entre categorías de términos y términos concretos. En concreto, en las ocasiones en la que no se ha utilizado una clasificación genérica, se genera un tipo de estructura de conocimiento diferente. Las diferencias fundamentales residen en que existe un número amplio de categorías generales válidas para ser gestionadas por una máquina, pero no para una persona. Por otra parte, esta clasificación de términos por categorías de palabras ha facilitado que la revisión pueda ser llevada a cabo por los ingenieros del dominio y que sólo en caso de duda o como resultado de esa clasificación un experto del dominio supervise el dominio. Tras diversas pruebas y estudios de clasificaciones similares, el número de categorías iniciales se ha establecido en torno a 15. La propuesta de una definición de una clasificación general para la construcción de sistemas de organización del conocimiento ha sido aplicada a proyectos enfocados a la construcción automatizada y a la construcción manual (p.e. a los distintos subdominios de SAGE-SP). Esta estructura permitía a los ingenieros del dominio incorporar vocabulario que había sido proporcionado por la compañía, en forma de pequeños listados. A medida que se confirmó el tipo de organización de la empresa se

precisaron los términos y se desecharon familias que no eran pertinentes para el dominio (por ejemplo, los gentilicios).

En el caso del dominio de REPSOL-YPF, se localizaron glosarios según las diferentes áreas que se querían modelar. Estos glosarios contenían términos propios del dominio y específicos, proporcionando un vocabulario normalizado. Por otra parte, se aportó la documentación propia de la empresa que a juicio de los expertos reflejaban suficientemente los dominios a modelar. Esta información fue entregada por temáticas que representaban cinco dominios diferentes, aunque con cierto grado de solapamiento.

En el caso de SAGE, la documentación que representaba el material primario para la construcción del sistema de organización del conocimiento consistía fundamentalmente en los ficheros de ayuda de los programas de sus aplicaciones informáticas. También se aportaban los ficheros de sugerencias y errores que se recogían de los clientes mediante el *call-center*.

La Guardia Civil aportaba para la construcción del tesoro la documentación que registran los miembros de la unidad para el seguimiento de los casos, donde se encontraba toda la información que se pretendía modelar, aunque en función de la investigación podían aparecer conceptos nuevos. El aumento del dominio era incremental debido fundamentalmente a la incorporación de nuevas instancias.

La Oficina del Defensor del Pueblo disponía de un recurso muy concreto y ya estructurado, su tesoro, con la información que se iba a tratar. Por otra parte, tenían a disposición de los expertos los informes que se tenían que indizar, permitiendo un adecuado reconocimiento y extracción conceptual.

Para el AGN, el proceso de especificación de los recursos documentales para la recopilación del material primario se determinó como necesario índices de topónimos, organigramas, clasificaciones internas, tipología documental, etc.

Desde el inicio de las experiencias con los proyectos se vieron las ventajas del uso de documentos estructurados o semiestructurados, en cuanto a la calidad y cantidad de conceptos y relaciones concentradas en este tipo documental, sin embargo en los entornos aplicados no han podido ser, prácticamente,

aprovechados como recursos documentales básicos, como se observa en la Tabla 1

La aplicación de un tratamiento especial aplicado a los documentos estructurados comenzó con la importación del tesauro utilizado por la Oficina del Defensor del Pueblo al gestor de tesauros *TmCake*. Este software permitió la evaluación y el mantenimiento del tesauro. REPSOL-YPF y SAGE-SP disponen y utilizan la versión actualizada de esta herramienta (actualmente denominada *Domain Reuser*), esta versión se encuentra más próxima a la metodología final propuesta.

En el caso del AGN, mediante la funcionalidad de exportación *del Domain Reuser* se reutilizaron partes de sistemas de organización del conocimiento como un tesauro toponímico de carácter general.

La extracción de conocimiento a partir de composición de palabras ha sido aplicada en las primeras fases de extracción de relaciones y organización de términos en el sistema de organización del conocimiento en los proyectos de REPSOL-YPF, SAGE y prototipo de la Guardia Civil. Incluso se puede aplicar cuando lo que se ha obtenido como recurso primario es una lista de términos simples y compuestos como listado de términos de indización.

El principal problema que supone este mecanismo es que se pueden establecer relaciones de generalización-especificación que no sean ciertas, porque el término que se considera específico sea un término compuesto lexicalizado que ha perdido la semántica del término que se establece como genérico. En los casos de REPSOL-YPF, SAGE, el prototipo para la Guardia Civil, así como las listas de términos que se utilizaron en el dominio del AGN, se realizó una revisión manual, para identificar posibles casos erróneos.

Por otra parte, es habitual que el tratamiento textual de recursos no estructurados produzca gran cantidad de términos que no son fáciles de discriminar como candidatos o no al KOS.

La experiencia de los procesos con REPSOL-YPF, SAGE, y el prototipo de la Guardia Civil sugieren que sea un proceso destinado a cerciorarse de relaciones que pueden ser afectadas por el contexto o la finalidad del sistema de organización del conocimiento. Con esto nos referimos por ejemplo a las relaciones de sinonimia o equivalencia, que aunque dos términos no sean sinónimos, podrían ser considerados como tal para un dominio concreto.

	Dominio	Familias	Conceptos	Media de Relaciones
REPSOL-YPF	<i>Medio-ambiente</i>	--	2224	1,08
	<i>Química</i>	--	3758	1,61
	<i>Refino</i>	--	4279	1,07
	<i>EyP</i>	--	2234	1,24
SAGE-SP	<i>Contabilidad</i>	15	3894	1,17
	<i>Nóminas</i>	21	2410	2,83
	<i>Facturación</i>	15	5584	1,08
GC	<i>Guardia Civil</i>	12	603	2,63

Tabla 2: Características de los KOS en diferentes entornos

En el desarrollo de los diferentes proyectos que se han realizado con las herramientas y las metodologías para la construcción automatizada de sistemas de organización del conocimiento, se observa que la definición y el desarrollo de los nuevos sistemas son más eficientes en cuanto a la extracción de relaciones. Otro dato a destacar es la calidad de las construcciones de los términos compuestos, y la descomposición de esas construcciones complejas. Sin duda, otra de las características que mejora considerablemente la adquisición de conocimiento es la extracción de entidades. Las mejoras afectan a la calidad de los términos específicos y a la especificidad de las relaciones.

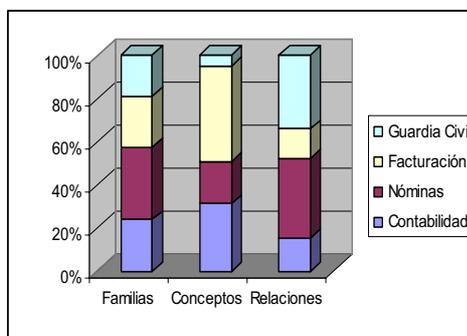
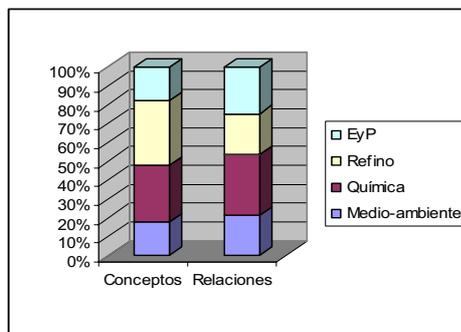


Figura 3: Visualización de las características de los KOS en los diferentes entornos

En el caso del prototipo para la Guardia Civil, el sistema de extracción de entidades tiene un impacto directo en la identificación de los términos candidatos a formar parte del sistema de organización del conocimiento, así como para la extracción de relaciones entre algunas de esas entidades. Otro impacto positivo que se refleja en los resultados es el aumento significativo de relaciones para el resultado de los sistemas de organización del conocimiento (Figura 3), debido a la fase de adquisición de relaciones mediante las unidades identificadas y a la flexibilidad del sistema de tratamiento textual e indizador del sistema de PLN.

4 Conclusiones

Esta propuesta se centra en mejorar los resultados con respecto a los aspectos más problemáticos de la construcción de KOS. Por una parte, en cuanto a las tareas asignadas a los expertos y responsables del dominio. Para ello se ha incidido en:

- minimizar el número de tareas asignadas
- reducir el tiempo de las tareas
- valorar y validar el conocimiento con información contextualizada
- formar a los expertos y responsables del dominio sobre el producto final
- mejorar la especificación de requisitos

El otro aspecto, en el que se ha centrado esta propuesta es en mejorar la calidad de los documentos que componen el corpus del dominio para la construcción de KOS mediante unos criterios para su construcción y la reutilización de recursos con vocabularios controlados existentes. Esta definición del corpus documental contribuye a:

- determinar los temas y facilitar las tareas del Ingeniero de Dominios para determinar los genéricos,
- determinar las expectativas del cliente,
- determinar un corpus de indización de mejor calidad y adaptado a sus necesidades

Por último, la mejora de las herramientas informáticas necesarias para la obtención de calidad resultados disminuye los errores de indización, extracción de información y construcción de KOS. Por lo tanto, cualquier

proceso de análisis y valoración tenderá a ser más efectivo y ha desempeñarse con mayor calidad. Asimismo un resultado de calidad favorecerá su uso, su utilidad y la necesidad de utilizar mecanismos de mantenimiento para la estructura de conocimiento. En este caso, las propuestas han estado orientadas a:

- diferenciar entre tipos de entidades
- corrección ortográfica para posibles deficiencias en los documentos
- organizar la extracción de términos y relaciones en distintas fases
- evaluación progresiva del conocimiento adquirido
- apoyo de una clasificación preexistente para la distribución del conocimiento

En resumen se ha propuesto un entorno para el desarrollo de KOS mediante una metodología configurable a diferentes escenarios. Para llevarla a cabo se debe elaborar de forma cuidadosa un corpus que contenga la información necesaria para la construcción del KOS y con unas aplicaciones específicas para la adquisición del conocimiento, y con un modelo de representación y construcción y mantenimiento del KOS.

Sin duda una de las ventajas logradas es la disminución de la dependencia de expertos del dominio, reduciendo los costes, las posibles inconsistencias entre expertos y la desmotivación que provocaban las tareas asignadas.

Bibliografía

- Aitchison, J.; Gilchrist, A.; Bawden, D. 1972. *Thesaurus construction and use: a practical manual*. 3rd ed. London: Aslib, 1997.
- Antoniou, G. y Harmelen, F. van. *A Semantic Web Primer*. London: The MIT Press, 2004.
- Baeza-Yates, R. y Ribeiro-Neto, B. *Modern Information Retrieval*. Massachusetts: Addison-Wesley, 1999.
- Berners-Lee, T.; Hendler, J.; Lassila, O.. The Semantic Web. *Scientific American Magazine*; May 2001
- Cabré Castellví, M^a. T. *La Terminología: Teoría, metodología y aplicaciones*. Barcelona: Antartida/Empuréis, 1993.
- Daconta, M. C.; obrst, Leo J. y Smith, K. T. *The Semantic Web. A guide to the future of XML, Web Services, and Knowledge Management*. Indianapolis: Wiley, 2003.

- Foskett, D. J. Thesaurus. *Encyclopaedia of Library and Information Science*. En: Spark-Jones, K. y Willett, P. (eds.). *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann, 1997. pp 111-134.
- Fraga, A.; Sánchez-Cuadrado, S. y Lloréns, J. Creación de un Tesoro Manual y Automático para el dominio de Arquitectura de Software. *Jornadas Chilenas de Computación, V Workshop Ingeniería del Software (WIS2005) de las Jornadas Chilenas de Computación*. Valdivia, Chile. 2005
- Gómez-Pérez, A.; Fernando-López, M.; Corcho, O. *Ontological Engineering*. London: Springer, 2004.
- Hodge, G. Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. The Digital Library Federation Council on Library and Information Resources. 2000
- Ingwersen, P. *Information Retrieval Interaction*. London: Taylor Graham, 1992 P. 245.
- ISO-2788: 1986. *Guidelines for the Establishment and Development of Monolingual Thesauri*. International Organization for Standardization, Second edition -11-15 UDC 025.48. Geneva: ISO, 1986.
- Lancaster, F. W. *Vocabulary control for information retrieval*. 2nd ed. Arlington, Virginia: Information Resources Press, 1986.
- Lloréns, J.. *Definición de una metodología y una estructura de repositorio orientadas a la reutilización: El tesoro de software*. Universidad Carlos III de Madrid, Departamento de Ingeniería, 1996
- Morato, J.; Lloréns, J.; Génova, G.; et al. Experiments in Discourse Analysis Impact on Information Classification and Retrieval Systems. *Information Processing and Management* 2003, 38. pp. 825-851.
- Novak, J. D. y D. B. Gowin, *Learning how to Learn*. New York: Cambridge University Press, 1984.
- Novak, J. D., *Learning, Creating , and Using Knowledge: Concept Maps as Facilitative Tools for Schools and Corporations*. Mahwah, N. J., Lawrence Erlbaum & Assoc, 1998
- Prieto-Díaz, R. Implementing Faceted Classification for Software Reuse. *Comm. ACM* 1991, 34 (5). pp. 88-97.
- Sánchez-Cuadrado, S. Definición de una metodología para la construcción automatizada de sistemas de organización del conocimiento. Tesis Doctoral. Universidad Carlos III de Madrid. Dpto. Biblioteconomía y Documentación, 2007.
- Sánchez-Cuadrado, S.; Lloréns, J, y Morato, J. Desarrollo de una aplicación para la gestión de relaciones en tesauros generados automáticamente. *Jotri 2003. II Jornadas de Tratamiento y Recuperación de la Información*. Madrid. 2003. pp. 151-156
- Sánchez-Cuadrado, S.; Lloréns, J. y Morato, J.; et al. Extracción Automática de Relaciones Semánticas. 2da. *Conferencia Iberoamericana en Sistemas, Cibernética e Informática. CISCI 2003*. Orlando, Florida. 2003a. pp. 265-268.
- Sánchez-Cuadrado, S.; y J. Morato Lara. Diseño de una herramienta para la Creación Asistida de KOS. VII Jornada de la Asociación Española de Terminología. Lenguas de especialidad y lenguajes documentales. 24 de noviembre de 2006.
- Van Slype, G.. *Los lenguajes de indización. Concepción, construcción y utilización en los sistemas documentales*. Madrid: Fundación Germán Sánchez Ruipérez. 1991.
- Zeng, M. L y L. Mai Chan. Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems. *Journal of the American Society for Information Science and Technology*, 2004, 55(5):377-395.