Fernanda Ribeiro e Maria Elisa Cerveira, org.

I Congresso ISKO Espanha e Portugal

XI Congreso ISKO España

7 a 9 de novembro de 2013

Informação e/ou Conhecimento: as duas faces de Jano

Atas

Faculdade de Letras da Universidade do Porto
CETAC.MEDIA
ISKO

Fernanda Ribeiro e Maria Elisa Cerveira, org.

I Congresso ISKO Espanha e Portugal

XI Congreso ISKO España

7 a 9 de novembro

Informação e/ou Conhecimento: as duas faces de Jano

Atas

Porto
Faculdade de Letras da Universidade do Porto
CETAC.MEDIA
2013

Ficha técnica:

Título: *Informação e/ou Conhecimento: as duas faces de Jano*

Autor: I Congresso ISKO Espanha e Portugal / XI Congreso ISKO España

Organização: Fernanda Ribeiro e Maria Elisa Cerveira

Edição: Faculdade de Letras da Universidade do Porto - CETAC.MEDIA

ISBN: 978-989-8648-10-5

Design e formatação: Ana Sofia Ramos

Apoios: Universidade do Porto / webQDA (Esfera Crítica Unip., Ld[a])

PROPUESTA DE ACTUALIZACIÓN DE MACRO-TESAUROS A PARTIR DE NOTICIAS DE DIVULGACIÓN CIENTÍFICO-TECNOLÓGICA

Updating proposal of macro-thesauri from popular science news

MARÍA-JOSÉ BAÑOS-MORENO
Universidad de Murcia
mbm41963@um.es

JUAN-ANTONIO PASTOR-SÁNCHEZ
Universidad de Murcia
pastor@um.es

RODRIGO MARTÍNEZ-BÉJAR
Universidad de Murcia
rodrigo@um.es

Resumen Los tesauros son herramientas de organización del conocimiento necesarias para el control de la información, más aún en el mundo de la información digital. Sin embargo, muchos de ellos adolecen de una falta de actualización que reduce considerablemente su utilidad. Este trabajo tiene precisamente como objetivo analizar el grado de actualización de dos de los tesauros más importante, el de la UNESCO y el de la Unión Europea (Eurovoc). Así, para la descripción de artículos de prensa, se procedió a la extracción de términos descriptivos de contenido a partir de titulares de artículos de divulgación científico-tecnológica publicados en prensa digital. A continuación mediante técnicas de Recuperación de Información se buscaron equivalencias con los tesauros mencionados anteriormente. Los resultados obtenidos muestran un nivel de equivalencia exacta o cercana que ronda el 50%. Este porcentaje se aproxima al 75% considerando equivalencias jerárquicas y asociativas. Este dato permite confirmar que ambos macro-tesauros pueden ser la base para elaborar otros vocabularios. En el caso concreto de su aplicación para la indización de noticias de divulgación científico-tecnológica, Eurovoc es ligeramente mejor que el Tesauro de la UNESCO, puesto que los términos y relaciones están más actualizados. El nivel de equivalencia exacta o cercana indica que las noticias de divulgación científico-tecnológica constituyen una fuente adecuada para la actualización de tesauros bien para la inclusión de nuevos términos o para la redefinición de las relaciones entre estos.

Palabras-clave Tesauro de la UNESCO. EUROVOC. Noticias de divulgación científica. Actualización de tesauros. Apache Solr.

Abstract Thesauri are knowledge organization instruments, necessary for the information control, especially in the digital information world. But many of them suffer a lack of update that reduces its usefulness. This work analyzes the updating level of two important thesauri: the UNESCO thesaurus and the European Union thesaurus (Eurovoc) to describe newspaper articles. We proceeded to the extraction of descriptive terms from popular science articles headlines published in digital press. Using Information Retrieval techniques equivalences were searched in both thesauri. Results show an exact equivalence level around 50%. Considering hierarchical and associative equivalences this percentage is close to 75%. This data confirms that both macro-thesauri can be the basis for developing other vocabularies. In case of their application for indexing popular science news, Eurovoc is slightly better than the UNESCO thesaurus, since terms and relationships are more updated. The exact or close equivalence level shows that popular science news are a proper source for updating thesauri for inclusion of new terms or to redefine the relationships between these.

Keywords UNESCO Thesaurus. EUROVOC Thesaurus. Popular scientific news. Updating thesaurus. Apache Solr.

Introducción

Con la generalización de Internet, buena parte de la actividad humana se ha desarrollado en la web, provocando un crecimiento exponencial de datos e información en la Red. El abaratamiento y desarrollo de las tecnologías de la información, la aparición de nuevas formas de conectividad y la ampliación de los espacios de almacenamiento (Sweeney, L. et al. 2001) no han hecho más que incidir en este aumento, generando "cantidades enormes de recursos desorganizados, duplicados o desactualizados, entre los que encontrar la información buscada termina resultando un trabajo arduo" (Pastor-Sánchez, J.A. 2009).

En este contexto, el control del vocabulario se revela como un procedimiento esencial (Soler-Monreal, C. y Gil-Leiva, I., 2011) que se materializa en el desarrollo de mecanismos para evitar la dispersión informativa y facilitar la localización de datos de manera rápida, precisa y relevante, aplicando herramientas como el tesauro (Observatorio Estatal de la Discapacidad, 2009). Este producto de organización del conocimiento (Smiraglia, R.P., 2012) es definido por la norma ISO 24964-1 (2011) como "un vocabulario estructurado y controlado en el que los conceptos son representados mediante términos," facilitando así la representación unívoca del contenido de documentos y consultas (Slype, G. van, 1991: 24).

El uso de un tesauro es prácticamente imposible sin una adaptación previa a la realidad de cada colección u organización en que se utiliza. Habitualmente se construye un tesauro *ad hoc* y desde cero, previo análisis de las necesidades específicas de descripción y recuperación de información del sistema, y mediante la incorporación de términos y relaciones obtenidos a partir de la indización de documentos del sistema (método deductivo) y/o de herramientas ya elaboradas, como diccionarios, enciclopedias, ontologías, manuales y tesauros (método inductivo).

Para ese proceso de inducción, dos de los macro-tesauros[1] más empleados son el de Unesco y el de Unión Europea o Eurovoc (en adelante "TUNESCO" y "TEUROVOC"). Una muestra de este uso concreto lo encontramos en Garrod, P. (2000), García Jiménez, A. (2002), del Valle Gastaminza, F. y García Jiménez, A. (2002), Shiri, A. et al. (2004), Kolar, M., et al. (2005), Castillo Blasco, L. (2006), Orenga-Gaya, L. y Giralt, O. (2011) y Fernández-Quijada, D. (2012), entre otros.

TUNESCO fue publicado por primera vez en 1977 por y para la propia organización, y se centra en los campos de educación cultura, comunicación e información, ciencias naturales, sociales y humanas (UNESCO, 1995). Desde entonces, ha sido actualizado varias veces (la última en 2008), traducido a múltiples idiomas (Ewketu, M. 2011; Fernández-Quijada D., 2012) y publicado en la web utilizando SKOS[2]. Se trata de un tesauro multidisciplinar, multilingüe y monojerárquico.

Por su parte, TEUROVOC fue creado en el seno de la Unión Europea para "gestionar con eficacia sus fondos [...] y permitir a los usuarios efectuar búsquedas documentales utilizando un lenguaje controlado". Ha sido actualizado en varias ocasiones (la última en 2012), traducido a todos los idiomas de la Unión y adaptado a la norma ISO 25964-1, a partir del primer borrador de la misma, publicado en 2009 (UNIÓN EUROPEA, 2012). Es multidisciplinar (aunque con un carácter político europeísta), multilingüe y polijerárquico.

Pero, para que un tesauro sea útil, es necesaria una actualización frecuente, con la revisión de términos y conceptos, y la redefinición de sus relaciones. Según Pérez Agüera, J.R. (2004) el objetivo no es otro que "incorporar la terminología derivada del desarrollo de la ciencia o materia a la que se dedica (...), cubrir lagunas o fallos detectados durante su utilización, así como adaptarlo a las necesidades de recuperación manifestadas por los usuarios". El mismo autor apunta precisamente que la falta de actualización es un problema habitual en estas herramientas, y lo es tanto desde un punto de vista conceptual (nuevos significados sin representación por uno o más significantes)

1 Un macro-tesauro, a diferencia de un tesauro, está "integrado por amplias áreas del conocimiento de las que cualquiera de ellas podrían dar lugar a un tesauro específico [denominado "micro-tesauro"] bien diferenciado". Por ejemplo, López Alonso, M.A. (2003), define cita el Macro-tesauro conceptual para los centros españoles de información juvenil.
2 Disponible en: http://skos.um.es/unescothes/?l=es

como léxico (nuevos sinónimos –significantes- no recogidos en la herramienta), lo que da lugar a tesauros desfasados y con una utilidad relativa, ya que son incapaces de representar adecuadamente el contenido de los documentos del sistema.

Este trabajo tiene como objetivos analizar el grado de actualización de TUNESCO y TEUROVOC y determinar su aplicación para describir artículos de prensa. También se plantea la posibilidad de renovar ambas herramientas a partir de palabras clave extraídas de textos periodísticos.

Para ello, tal como se describe en el apartado de metodología, se han indizado titulares de artículos divulgación científico-tecnológica de periódicos de los países más representativos en estas áreas. Posteriormente se detalla el proceso por el que Apache Solr ha buscado equivalencias entre los términos extraídos y los de ambos tesauros. A continuación se discute acerca de los resultados obtenidos en relación a determinados aspectos positivos y negativos de los macro-tesauros en cuanto al nivel de actualización de sus términos y relaciones. Finalmente, se presenta una serie de conclusiones y sugerencias a partir de los resultados obtenidos.

1 Metodología

Para la actualización de tesauros, como hemos visto, podemos emplear un método deductivo, extrayendo términos y relaciones de documentos internos del sistema. En este caso, por el contrario, hemos recurrido a elementos externos, esto es, titulares de noticias publicadas en diarios internacionales, como fuente para esa extracción, ya que:

1) Se trabaja sobre un corpus de menor tamaño. Las noticias tienen una extensión menor, en general, que cualquier informe o documento de trabajo integrado en un sistema documental como el de la UNESCO u EUROVOC, por ejemplo, lo que agiliza el análisis de contenido;

2) La información periodística posee unas cualidades que la convierten en una fuente de enorme potencial para la renovación de muchos tesauros, sean de ámbito general o especializado. Concretamente, en el caso de TUNESCO y TEUROVOC, destacan las siguientes:

   - Actualidad: Buena parte de lo que ocurre en el mundo se refleja y transmite al público a través de la información periodística (sea por medios convencionales o no). Esto nos asegura que el tesauro incorporará los términos más utilizados y/o novedosos;
   - Inmediatez, especialmente a través de Internet. El tiempo que transcurre entre un acontecimiento y su publicación en prensa es reducido (y cada vez más);
   - Cobertura geográfica y temporal: Un asunto con carácter noticioso es analizado por medios locales e internacionales a cuya información se puede acceder vía web en cualquier momento y lugar;
   - Contrastación de la información antes de su publicación. Así, las noticias están avaladas por la política de trabajo del medio. Además, la cobertura simultánea de un mismo hecho facilita un conocimiento más completo de lo acontecido;

- Normalización. A través de las guías de estilo de cada diario se establece, entre otros, cómo se transcribirán determinados neologismos, palabras en otros idiomas, etc;
- Generalidad y especialización. Los diarios recogen noticias de ámbito general y especializadas. En este último caso suelen ser redactadas por periodistas que, habitualmente, tienen una formación en el área o, directamente, especialistas cuya actividad profesional se circunscribe en la misma.

3) La sección de divulgación científico-tecnológica poseen además otras condiciones propias de su especificidad:

- Interés divulgativo. La información periodística informa, forma y entretiene, cumpliendo "un papel fundamental en la divulgación de los principales descubrimientos del siglo XX" y XXI (Fernández Muerza, A., 2005a y 2005b);
- Empleo de fuentes acreditadas: Las noticias de divulgación se nutren de revistas científicas para sus publicaciones, con información estable y sometida a una potente revisión y, por tanto menos susceptible a modas, tendencias y recursos lingüísticos;
- Uso y adaptación de lenguajes técnicos (Castillo Blasco, L. 2006) para "recontextualizar aspectos del conocimiento o de la práctica científica" (Alcíbar Cuello, M., 2004).

Teniendo en cuenta lo anterior se procedió a diseñar una muestra seleccionando sucesivamente países, periódicos y titulares de noticias. En primer lugar se estableció un ranking de países punteros en ciencia y tecnología, a partir de cinco parámetros previamente definidos, que miden el desarrollo científico-tecnológico, seleccionando aquellas naciones que se situaban en los primeros puestos de, al menos, tres de estos cinco parámetros:

- Gasto Interior Bruto en I+D+i (2008).
- Valor bruto (en dólares) de exportaciones de alta tecnología (2008).
- Valor bruto (en dólares) recibidos por cada país, en concepto de royaltys, cánones y otros por uso de patentes (2008).
- Nº de artículos publicados por país (2008).[3]
- Total de premios recibidos, por país, más relevantes en Ciencia y Tecnología. [4]

Posteriormente, a partir de los datos ofrecidos por *4International Media & Newspaper*[5], se seleccionó el periódico de información general de ámbito nacional más popular en cada país. Esto es, el diario más leído teniendo en cuenta datos de lectura y acceso a la edición digital e impresa (dos en el caso de Estados Unidos y China). Las noticias se seleccionaron de las secciones de Ciencia

3 Los datos de estos cuatro primeros parámetros proceden de Banco Mundial (2012) del año 2008.
4 Premios analizados: Premio Nobel, Medalla Internacional para Descubrimientos Sobresalientes en Matemáticas, Premios Príncipe de Asturias, Premio Abel, Premios Albert-Einstein, Medalla Wollaston, Premio Mundial de Tecnología, Premio Turing y Premio Kyoto.
5 Disponible en: http://www.4imn.com

y/o Tecnología de cada medio y fueron recogidas, durante cuatro meses, comprendidos entre el 9 de marzo y 9 de julio de 2012.

A continuación se realizó una indización asistida por ordenador y en lenguaje natural, mediante la que se extrajeron entre 1 y 6 palabras clave de los titulares de noticias y se tradujeron a tres idiomas: Español, Francés e Inglés con el fin de determinar la correspondencia terminológica entre los conceptos a los que se hacían referencia en los titulares. Esta decisión se tomó para disponer de un mecanismo que permitiera el uso cruzado de dichos idiomas como lenguajes pivote para desambiguar los casos de homonimia (Areas da Luz Fontes et. al, 2010; Degani & Tokowicz, 2010; Liang, 2001). Para ello se utilizaron herramientas de uso común, como *Word Reference*, *Linguee*, *Google Translator* y *Wikipedia*.

Desde el punto de vista metodológico se estructuró el proceso considerando los siguientes factores que, según Nakurawa, M.C. (2009) influyen en la indización:

- **Qué se indiza.** Titulares de noticias de divulgación científica publicadas en diarios de información general. Cuando el titular no es suficientemente descriptivo, la noticia se descarta, salvo que haya sido cubierta por otro diario ya analizado (y por tanto, conocida);
- **Quién indiza.** La formación, la experiencia en indización, el conocimiento del asunto y el dominio de la herramienta de indización son factores clave. En este caso, ya se había trabajado previamente con ambos macro-tesauros;
- **En qué contexto se analiza.** Política de indización: como norma general, siguiendo a Currás, E. (1991), se utiliza la forma singular; salvo en ciertos casos, donde es más práctico elegir la forma plural para dar mayor sentido a los términos y evitar ambigüedades y confusiones; necesidades de los usuarios; carga de trabajo; tiempo dedicado. En este caso se decidió decida entre 2 y 3 minutos a la extracción de palabras clave por noticia.

Más adelante se procedió a realizar una búsqueda de equivalencias de los términos extraídos con respecto a ambos tesauros. Para ello se construyeron sendas colecciones de documentos a partir de la terminología de ambos herramientas, que fueron indizadas en el sistema de recuperación de información Apache Solr. De esta forma en un único documento se agruparon (en campos separados) tanto los descriptores como los no-descriptores de cada concepto en Español, Inglés y Francés. Después se buscaron automáticamente las equivalencias entre los términos extraídos de los titulares de prensa y los términos de ambos tesauros. Finalmente se procedió a evaluar los resultados obtenidos, determinando el grado de equivalencia entre los conjuntos anteriores.

2 Resultados

Siguiendo los criterios indicados en la metodología, se conformó un corpus de 1599 noticias, de las que para este trabajo se seleccionaron 320 titulares (20%), tal y como se indica en la tabla 1:

País	Medio	Secciones CyT	Nº artículos
Alemania	Süddeutsche Zeitung	Wissen	19
Canadá	The Global and Mail	Technology	21
China	China Daily	Sci-Tech	13
	The China Post (Taiwán)	Life Health	21
República de Corea	The Korea Times	Science Technology	22
España	El Mundo	Ciencia	19
Estados Unidos	The New York Times	Science	41
	The Washington Post	Energy & Environment Health & Science Innovations	38
Francia	Le Monde	Planète Technologies	41
Italia	La Reppublica	Scienze Tecnologia Ambiente	5
Japón	Yomiuri Shimbun	Science & Nature	40
Reino Unido	The Daily Telegraph	Science Tecnology	11
Rusia	Pravda	Science	29
		Total artículos	**320**

Tabla 1. Diarios seleccionados para la extracción de noticias

Tras la indización, se obtuvieron 1018 palabras clave, que se redujeron a 599 al eliminar las duplicaciones. A continuación, se tradujeron a castellano, inglés y francés (ver ejemplo en Tabla 2), obteniendo así un corpus de 1.797 términos.

CASTELLANO	INGLÉS	FRANCÉS
Acuicultura	Aquaculture	Aquaculture
Brecha digital	Digital divide	Fossé numérique
Consumo de agua	Water consumption	Consommation d'eau
Economía verde	Green economy	Économy verte
Periodismo participativo	Participative Journalism	Journalisme participatif
Prevención del crimen	Crime prevention	Prévention du crime
Tecnología inalámbrica	Wireless technology	Technologie sans fil
Vandalismo	Vandalism	Vandalisme

Tabla 2. Muestra de términos extraídos en castellano, inglés y francés

Después se construyeron sendas colecciones a partir de los términos de cada tesauro, sobre las que se efectúan las interrogaciones a partir de los términos extraídos. La estructura para el almacenamiento de cada concepto fue la siguiente:

NOMBRE DEL CAMPO	DESCRIPCIÓN
id	Identificador del concepto
type	Tesauro del concepto (eurovoc ó unesco)
des_es	Término descriptor en Español
des_fr	Término descriptor en Francés
des_en	Término descriptor en Inglés
nd_es	Término no-descriptor en Español
nd_fr	Término no-descriptor en Francés
nd_en	Término no-descriptor en Inglés

Tabla 3. Estructura de campos de los conceptos para su indización como documentos en Apache Solr

Después se buscaron automáticamente las equivalencias entre los términos extraídos de titulares de prensa y los términos de ambos tesauros. Se realizaron una serie de búsquedas para cada término e idioma. Tras diversos ensayos se obtuvo un procedimiento en el que se realizaban siete búsquedas (por término e idioma):

- Búsqueda por palabras en índice general (Q1).
- Búsqueda literal en el campo descriptor (Q2).
- Búsqueda literal en el campo no-descriptor (Q3).
- Búsqueda lematizada de expresión en el campo descriptor (Q4).
- Búsqueda lematizada de expresión en el campo no-descriptor (Q5).
- Búsqueda lematizada por palabras en campo descriptor (Q6).
- Búsqueda lematizada por palabras en campo no-descriptor (Q7).

Apache Solr proporciona un medida de similitud o *score*[6] de la consulta con cada uno de los documentos del sistema. Resulta evidente que las equivalencias literales exactas de los términos de los titulares con descriptores y no descriptores permiten determinar una identificación de conceptos exacta o muy cercana. Por este motivo a los resultados obtenidos en las consultas Q2 y Q3 se les ha aplicado un factor de potenciación de la medida de similitud o *boost* de 5 y 3 respectivamente. Experimentalmente también se comprobó la necesidad de potenciar los resultados de las búsquedas por palabras en el índice general (consulta Q1), aplicando en este caso un *boost* de 2,5.

Los primeros datos obtenidos al aplicar esta técnica aconsejaron establecer para la búsqueda general y las lematizadas un umbral mínimo de score debajo del cual debían desecharse dichos resultados.

Los resultados alcanzados se analizaron para determinar el tipo de equivalencia existente entre los términos de los titulares y los recogidos en los tesauros y se clasificaron de la siguiente forma:

- **Término correcto (TC)**: Los términos de los titulares y del tesauro coinciden tanto en significante como en significado. Por ejemplo el término "estudiante" extraído de un titular tiene una equivalencia TC con el término "estudiante" de un tesauro;

[6] Esta medida está basada en el método TF-IDF. Más información en: https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html

- **Término sinónimo (TS)**: Los significantes de titulares y tesauro son diferentes pero hacen referencia al mismo significado. Es el caso de "acuerdo empresarial" que tiene una equivalencia TS de "acuerdo interempresarial";
- **Término específico (TE)**: El término del titular representa un concepto más específico que el del término en el tesauro. Ejemplo: "célula solar" es TE de "colector solar";
- **Término genérico (TG)**: El significante del titular representa un significado más genérico que el recogido por el tesauro. Incluye relaciones de clase y partitivas. Las relaciones enumerativas, a excepción de las de algunos organismos internacionales y geográficas, se reemplazaron por un término genérico. Por ejemplo, "estudio" tiene una equivalencia TG de "estudio de impacto";
- **Término relacionado (TR)**: Los significados representados por los significantes extraídos y los recogidos en el tesauro mantienen cierta relación semántica entre sí: "ladrón de bancos" es TR de "banco";
- **Término nuevo (TN)**: No se encuentra relación entre significante extraído y significantes en tesauros. Ejemplo: "smartphone" es un TN, no recogido en TEUROVOC;
- **Término con falsa equivalencia (TFE)**: El término del titular no mantiene relación con el devuelto por Apache Solr, constituyendo, por tanto, una relación de falsa equivalencia como ocurre con "agujero negro" es TNM de "mar negro".

No todos los tipos de equivalencia tienen el mismo nivel de relevancia. Por este motivo es posible afirmar que siempre es preferible una equivalencia exacta (TC) a una de sinonimia (TS). También es posible que es preferible una equivalencia de TE a una TG, puesto que en el caso de una equivalencia específica el significado del término de un titular es cubierto por el del tesauro, lo que no sucede al contrario.

	TEUROVOC		**TUNESCO**	
	Total de equivalencias	**Porcentaje de equivalencias**	**Total de equivalencias**	**Porcentaje de equivalencias**
TC	290	48,41%	264	44,07%
TS	10	1,67%	14	2,34%
TG	43	7,18%	25	4,17%
TE	42	7,01%	69	11,52%
TR	70	11,69%	75	12,52%
TFE	83	13,86%	69	11,52%
TN	61	10,18%	83	13,86%

Tabla 4. Resultados obtenidos con Solr para los términos de los tesauros

Respecto a las equivalencias con TEUROVOC, destaca lo siguiente: para casi la mitad de los términos (48,41%) se encontró una equivalencia exacta en el tesauro. Más de un 14% de los términos extraídos guardan una relación de jerarquía con los términos del tesauro y casi un 12% mantienen una relación asociativa. Para algo más de un 24% (144 términos) no se halló ningún tipo de relación o fue una falsa equivalencia, por lo que fueron excluidos del análisis.

Referente a TUNESCO: Poco más del 44% de los términos tenían una equivalencia exacta en el tesauro. Casi un 16% de las palabras clave obtenidas mantienen una relación de jerarquía con los términos del tesauro. Poco más de un 12% de los términos guardan una relación de tipo asociativo.

Además, se identificaron más de un 25% de términos extraídos para los que Apache Solr no halló relación o era falsa, por lo que, como en el caso de TEUROVOC se excluyeron del análisis.

A partir de los datos anteriores y para evaluar la eficacia del procedimiento seguido, se calculó la precisión *P* de la búsqueda efectuada en cada macro-tesauro [(Cleverdon, C.W. et al. (1966), Tolosa, G.H. y Bordignon, F. (2008) y Hage, W.R. van et al. (2010)]. Considerando TR_{REL} y TR_{TOT} como términos relevantes recuperados y total de términos recuperados, respectivamente, se tendría la siguiente ecuación:

$$Precisión(P) = \frac{TR_{REL}}{TR_{TOT}}$$

Es posible realizar varios cálculos de precisión: la precisión exacta P_{EXACT} consideraría como relevantes únicamente los términos correctos; la precisión cercana P_{CLOSE} también tendría en cuenta los términos sinónimos; la precisión total P_{TOTAL} incluiría cualquier tipo de relación. De este modo se obtendría los siguientes valores de precisión:

	TEUROVOC	**TUNESCO**
P_{EX}	0,4841	0,4407
P_{CLOSE}	0,5008	0,4641
P_{TOT}	0,7595	0,7462

Tabla 5 Resultados de precisión exacta, cercana y total obtenidos con Solr

3 Discusión

Alrededor de 3/4 de las palabras clave extraídas guardan algún tipo de relación tanto con los términos de TEUROVOC (75,96%) como con los de TUNESCO (74,62%), lo que da lugar a una precisión relativa P_{TOTAL} alta en ambos casos, situada en torno al 0,75. Esto indica que Solr es capaz de detectar la existencia de relaciones entre palabras clave extraídas de los titulares y términos presentes en ambos macro-tesauros, aún cuando éstas no están recogidas como tal en los vocabularios. De estas equivalencias más de un 25% son de tipo jerárquico o asociativo. Es decir, una búsqueda eventual por dichas palabras clave no aportará el resultado más adecuado, produciendo silencio o ruido, dependiendo del caso.

En este sentido, ambos tesauros revelan cierta falta de revisión de sus términos y relaciones, al menos en los ámbitos que nos ocupan. Quizá la especificidad de las áreas analizadas, Ciencia y Tecnología, provoca un incremento en los porcentajes de relaciones jerárquicas y asociativas, aunque bien es cierto que TEUROVOC incluye una amplia terminología tecnológica, derivada de la profusa normativa desarrollada al efecto en la Unión Europea. Por otro lado, TUNESCO recoge expresiones más propias del ámbito de la Ciencia y la Educación, lo que puede dar lugar a lagunas en la parte tecnológica que puede que hayan motivado estos datos.

Sin embargo, aproximadamente la mitad de los términos extraídos, clasificados como TN y TS, de acuerdo a la búsqueda efectuada por Apache Solr, son recogidos por los dos tesauros. Por esta razón, las precisiones exacta y cercana, (P_{EXACT} y P_{CLOSE}) son bastante más pequeñas que la precisión total. De esta forma, al indizar noticias de divulgación científico-tecnológica, más de la mitad de los términos empleados, en ambos casos, encuentran una equivalencia exacta o cuasi-exacta en ambos

tesauros. Ello nos indicaría que, pese a lo señalado en el párrafo anterior, existe cierta preocupación en ambas instituciones por actualizar estos vocabularios.

Por otro lado, para casi un 25% de los términos extraídos no se encontró ningún término con el que relacionar en ambos vocabularios. En unos casos, se produce ruido, cuando Apache Solr muestra términos con los que, en realidad, no existe relación (falsos equivalentes). En otros, se genera silencio, cuando el programa no es capaz de encontrar ninguna relación satisfactoria con los términos recogidos en los macro-tesauros. Aquí se encuentra el auténtico *filón* para la actualización de estos vocabularios, pues que una cuarta parte de los términos extraídos no aparezcan representados debe llevar, cuando menos, a plantearse la utilidad de las noticias de secciones especializadas para dicha actividad, más aún si tenemos en cuenta las características de las noticias de divulgación científico-tecnológico señaladas anteriormente. En determinados casos, se podría plantear la inadecuación de los términos extraídos de las noticias teniendo en cuenta la finalidad y características de cada macro-tesauro. Así "macho alfa" podría ser recogido en TUNESCO, por estar más enfocado a Ciencia, pero no en TEUROVOC.

Conclusiones y líneas futuras

Los datos obtenidos permiten confirmar que ambos macro-tesauros pueden utilizarse como base para la elaboración de otras herramientas de gestión del conocimiento, siendo necesaria una adaptación a las necesidades de sus usuarios. En el caso concreto de indización de noticias de divulgación científico-tecnológica con estos vocabularios o con herramientas creadas a partir de los mismos, podría decirse que TEUROVOC es ligeramente más adecuado: aunque TUNESCO tiene un carácter más enciclopédico y universal, la frecuencia de revisión de TEUROVOC es mayor, haciendo que los términos y relaciones estén más actualizados.

Las noticias de divulgación científico-tecnológica son una fuente adecuada para la actualización de tesauros ya que, como se ha visto, cerca de la mitad de las palabras clave extraídas de las noticias son recogidas por los macro-tesauros pero el resto, no, aunque buena parte de éstos sí guardan una estrecha relación de jerarquía o asociación con los de los vocabularios. De esta forma, este tipo de información puede utilizarse bien para la inclusión de nuevos términos no contemplados hasta ahora en los vocabularios o para la redefinición de las relaciones entre los ya recogidos.

La especificidad de la temática de los artículos periodísticos utilizados, procedentes de las secciones de Ciencia y Tecnología o similares facilita la renovación de micro-tesauros concretos de estas áreas. La actualización de otros tesauros especializados y/o micro-tesauros podría llevarse a cabo también utilizando noticias de prensa de otras secciones, por ejemplo, economía o cultura. Para ello, se debe calcular previamente la precisión de los términos recogidos en los tesauros con respecto de los extraídos mediante la indización de las noticias de estas secciones.

Asimismo, las noticias de prensa podrían utilizarse para desarrollar y/o actualizar otro tipo de herramientas de gestión del conocimiento, más allá de los tesauros. Tal sería el caso de las ontologías, donde las relaciones entre conceptos son mucho más complejas y definidas.

Referencias

ALCÍBAR CUELLO, M. (2004). La divulgación mediática de la ciencia y la tecnología como recontextualización discursiva. *Anàlisi: Quaderns de comunicació i cultura*, (31), 43–70.

AREAS DA LUZ FONTES, A. B.; YEH, L. H.; SCHWARTZ, A. I. (2010). Desambiguação lexical bilíngue: a natureza dos efeitos de coativação lexical entre as línguas. *Letrônica: Revista Digital do PPGL*, 3(1). Recuperado de http://revistaseletronicas.pucrs.br/ojs/index.php/letronica/article/view/7074.

BANCO MUNDIAL. (2012). *Indicadores del desarrollo mundial: PIB (US$ a precios actuales) (Estadística).* Recuperado de http://datos.bancomundial.org/indicador/NY.GDP.MKTP.CD.

CASTILLO BLASCO, L. (2006). *Elaboración de un tesauro de información de actualidad y conversión en red semántica para su empleo en un sistema de recuperación periodístico*. Universidad de Valencia, Valencia. Recuperado de http://www.tdx.cat/bitstream/handle/10803/9982/castillo.pdf?sequence=1.

CLEVERDON, C. W.; KEEN, M. (1966). *Aslib Cranfield research project - Factors determining the performance of indexing systems;* Volume 2, Test results (p. 299). Crandfield: National Science Foundation. Recuperado de http://dspace.lib.cranfield.ac.uk/handle/1826/863.

CURRÁS, E. (1991). *Thesauros. Lenguajes terminológicos*. Madrid: Paraninfo.

DEGANI, T.; TOKOWICZ, N. (2010). Semantic ambiguity within and across languages: An integrative review. *The Quarterly Journal of Experimental Psychology*, 63(7), 1266-1303.

DEL VALLE GASTAMINZA, F. del; GARCÍA JIMÉNEZ, A. G. (2002). Construcción de un tesauro para el Centro de Documentación de Telecinco. *Scire: Representación y organización del conocimiento*, *8*(1), 103–113.

DIALNET SNAPSHOT. (n.d.). Recuperado de http://dialnet.unirioja.es/servlet/articulo?codigo=1301998.

EWKETU, M. (2011, November 28). *The UNESCO Thesaurus*. Presented at the UN-LINKS Meeting, París. Recuperado de http://www.unesco.org/library/PDF/The%20UNESCO%20Thesaurus.pdf.

FERNÁNDEZ MUERZA, A. (2005a). *Estudio del periodismo de información científica en la prensa de referencia: el caso español a partir del análisis comparativo*. Universidad del País Vasco. Recuperado de http://e-ciencia.com/afm/tesis-alex.pdf.

FERNÁNDEZ MUERZA, A. (2005b). La información científica en la prensa de referencia: el caso español a partir de un análisis comparativo. *ZER Revista de Estudios de Comunicación*, *19*, 205–232.

FERNÁNDEZ-QUIJADA, D. (2012). El uso de tesauros para el análisis temático de la producción científica: apuntes metodológicos desde una experiencia práctica. *BiD: textos universitaris de biblioteconomia i documentació*, *29*. Recuperado de http://www.ub.edu/bid/29/fernandez2.htm.

GARCÍA JIMÉNEZ, A. G. (2002). *Metodología de validación del análisis documental y de los lenguajes documentales en el discurso periodístico*. Universidad Complutense de Madrid, Madrid. Recuperado de http://www.ucm.es/BUCM/tesis/19911996/S/3/S3005101.pdf.

GARROD, P. (2000). Use of the "UNESCO Thesaurus" for archival subject indexing at UK-NDAD (UK-National-Digital-Archive-of-Datasets, database, terms, web, online catalogues). *Journal of the Society of Archivists*, *21*(1), 37–54. doi:10.1080/00379810050006902.

GIL-LEIVA, I. (2007). The indexing at the Internet. *Brazilian Journal of Information Science*, (2), 47–68.

INTERNATIONAL MONETARY FUND. (2013). *World economic outlook : a survey by the staff of the International Monetary Fund*. Washington, DC: International Monetary Fund. Recuperado de http://www.zotero.org/styles/iso690-numeric-en

ISO. (2011). ISO 25964-2:2011. *Thesauri and interoperability with other vocabularies*. Part 1: Thesauri for information retrieval.

KOLAR, M.; VUKMIROVIC, I.; BASIC, B. D.; SNAJDER, J. (2005). *Computer aided document indexing system*. (V. L. Luzar & V. H. Dobric, Eds.). Zagreb: Srce Univ Computing Centre, Univ Zagreb.

LIANG, G. B. M. J. (2001). Experiments in Trilingual Cross-Language Information Retrieval. In *Proceedings 2001 Symposium on Document Image Understanding Technology* (Univ. Maryland), 169-179.

LÓPEZ ALONSO, M. Á. (2003). Compilación de un macro-tesauro conceptual para los centros españoles de información juvenil. *Scire: Representación y organización del conocimiento*, *9*(1), 47–56.

NAGYPÁL, G. (2005). Improving information retrieval effectiveness by using domain knowledge stored in ontologies. In *On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops* (Vol. 3762, pp. 780–789). Presented at the OTM Confederated International Workshops and Posters, Cyprus: Springer. Recuperado de http://dip.semanticweb.org/documents/Gabor-Nagypal-Improving-information-retrieval-effectiveness-by-using-domain-knowledge-stored.pdf.

NARUKAWA, C. M.; LEIVA, I. G.; FUJITA, M. S. L. (2009). Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software SISA com uso da terminologia DeCS na área de Odontologia. *Informação & Sociedade: Estudos*, *19*(2). Recuperado de http://www.ies.ufpb.br/ojs/index.php/ies/article/view/2925.

OBSERVATORIO ESTATAL DE DISCAPACIDAD. (2009). *Tesauro de la discapacidad*. Badajoz: Observatorio Estatal de Discapacidad. Recuperado de http://www.observatoriodeladiscapacidad.es/sites/default/files/tesauro%20de%20la%20discapacidad.pdf.

ORENGA-GAYA, L.; GIRALT, O. (2011). The official gazette of the Generalitat de Catalunya: genesis of a digital newspaper. *Profesional De La Informacion*, *20*(3), 340–344. doi:10.3145/epi.2011.may.14.

PASTOR-SÁNCHEZ, J.-A. (2009). *Diseño de un sistema colaborativo para la creación y gestión de tesauros en Internet basado en SKOS*. Universidad de Murcia, Murcia. Recuperado de http://www.tesisenred.net/handle/10803/10914.

PÉREZ AGÜERA, J. R. (2004). Automatización de tesauros y su utilización en la web semántica. *BiD: textos universitaris de biblioteconomia i documentació*, *13*. Recuperado de http://www.ub.edu/bid/13perez2.htm.

SHIRI, A.; NICHOLSON, D.; MCCULLOCH, E. (2004). User evaluation of a pilot terminologies server for a distributed multi-scheme environment. *Online Information Review*, *28*(4), 273–283. doi:10.1108/14684520410553769.

SLYPE, G. van. (1991). *Los lenguajes de indización.* Concepción, construcción y utilización en los sistemas documentales. Madrid: Pirámide.

SMIRAGLIA, R. P. (2012). Organización del conocimiento: Algunas tendencias en un dominio emergente. *El Profesional de la Información*, *21*(3), 225–227.

SOLER MONREAL, C.; GIL LEIVA, I. (2011). Evaluation of controlled vocabularies by inter-indexer consistency. *Information Research*, *16*(4). Recuperado de http://informationr.net/ir/16-4/paper502.html.

SWEENEY, L. (2001). Information Explosion. In L. Zayatz, P. Doyle, J. Theeuwes; J. Lane (Eds.), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Washington, DC: Urban Institute. Recuperado de http://dataprivacylab.org/dataprivacy/projects/explosion/explosion2.pdf.

TOLOSA, G. H.; BORDIGNON, F. R. A. (2008). *Introducción a la Recuperación de Información: Conceptos, modelos y algoritmos básicos*. Buenos Aires: Universidad Nacional de Luján. Recuperado de http://hdl.handle.net/10760/12243.

UNESCO. (1995). Tesauro de la Unesco. *Construir la paz en la mente de los hombres y de las mujeres*. Recuperado February 14, 2013, de http://databases.unesco.org/thessp/.

UNIÓN EUROPEA. (2012). *EuroVoc, tesauro multilingüe de la Unión Europea*. Recuperado May 9, 2013, de http://eurovoc.europa.eu/drupal/?q=es.

VAN HAGE, W. R.; SINI, M.; FINCH, L.; KOLB, H.; SCHREIBER, G. (2010). The OAEI food task: An analysis of a thesaurus alignment task. *Appl. Ontol.*, *5*(1), 1–28.