Fuentes para la actualización de macro-tesauros: noticias de divulgación científica

Sources for updating macrothesauri: popular science news

María-José Baños-Moreno

Universidad de Murcia, 30100, mbm41963@um.es

Resumen

Los tesauros son herramientas de organización del conocimiento necesarias para el control de la información. Muchos no se actualizan con la necesaria frecuencia, reduciendo considerablemente su utilidad. Este trabajo tiene como objetivos, analizar el grado de actualización de los tesauros de la UNESCO y Unión Europea (EUROVOC); y determinar la capacidad de las noticias de divulgación científica de ser usadas para efectuar esa renovación. Para ello se han extraído palabras clave a partir de titulares de artículos de divulgación científica, publicados en prensa digital. A continuación, mediante técnicas de Recuperación de Información (N-gramas y Apache Solr) se buscaron equivalencias con los tesauros mencionados anteriormente. Los datos obtenidos permiten confirmar que ninguno de los macro-tesauros debe utilizarse para describir noticias de divulgación científica, ya que buena parte de las palabras clave extraídas no se recogían en estos. Sin embargo sí pueden emplearse como base para la construcción de otros vocabularios de organización del conocimiento. Por otro lado, una revisión más frecuente es necesaria y obligatoria. La inclusión de nuevos términos y la redefinición de las relaciones entre los ya existentes incrementarían incuestionablemente la calidad de las herramientas analizadas. Las noticias de divulgación resultan adecuadas para dicha actualización, constituyendo auténticos yacimientos de conceptos y términos.

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

Abstract

Thesauri are knowledge organization tools necessary to control information, but many of them are not updating with the frequency required, it reduces their usefulness. This paper aims to analyze the updating degree of UNESCO thesaurus and European Union thesaurus (EUROVOC) and determine if popular science news are useful for this renovation. In this work, keywords were extracted from headlines. After, using Information Retrieval techniques (N-grams and Apache Solr) they were searched in both thesauri to find equivalences. Data obtained let us confirm that none of these macro-thesauri should be used to describe popular science news, although both of them can be used as a base for building other knowledge organization vocabularies. On the other hand, a more frequent review is necessary. Inclusion of new terms and redefinition of relationships between them is unquestionable to increase the quality of analyzed tools. Popular science news are adequate to this purpose, they constitute real sources of concepts and terms.

Keywords: UNESCO Thesaurus, EUROVOC Thesaurus, Updating thesauri, N-gramas method, Popular science news.

NOTA DEL EDITOR: este trabajo se basa en la comunicación "Propuesta de actualización de macro-tesauros a partir de noticias de divulgación científico-tecnológica" (Baños-Moreno, Pastor-Sánchez, Martínez-Béjar, en prensa) presentada al I Congreso ISKO España y Portugal / XI Congreso ISKO España (2013), cuyas actas aún no han sido publicadas.

1 Introducción.

El ritmo exponencial del crecimiento de la información (Carrizo Sainero, 2000, p.1-5, Rodríguez Mateos, 2011, p. 90) y su dispersión son unos de los grandes problemas de la sociedad del conocimiento, que disciplinas como la Informática y la Documentación ofrecen desde diferentes puntos de vista, y que desembocan en el desarrollo de productos concretos como los tesauros, cuya finalidad es doble: controlar la información (saber qué hay, dónde está, cuáles son sus características y qué contiene) y facilitar su recuperación (quién la necesita, cómo la necesita y cómo accede). Pero los tesauros son útiles no sólo para la organización y recuperación de documentos, también como fuente para la creación de otros vocabularios (folksonomías, nubes de etiquetas, taxonomías, ontologías, etc.), Desde esta consideración, el Tesauro de la UNESCO y

el de la Unión Europea (Tunesco y Teurovoc, en adelante) constituyen piezas fundamentales para la generación de otros productos a partir de la reutilización de sus términos, relaciones y notas. Así lo confirman los múltiples trabajos centrados en su análisis y aplicación: Pollit et al. (1995), Garrod (2000), Baxter et al. (2002), Pouliquen et al. (2004), Shiri et al. (2004), Kolar et al. (2005), Saric et al. (2005), Smits & Friis-Christensen (2007), Fiser & Sagor (2008), Mastora et al. (2008), Loza Mencía & Fuernkranz (2008), Campos & Romero (2009), Daudaravicius (2010), Shvaiko et al. (2010), Orenga-Gaya & Giralt (2011), Martínez et al. (2011) y Fernández-Quijada (2012), entre otros.

En la siguiente tabla se muestran las características generales de cada herramienta.

	Tunesco	*Teurovoc*
Entidad	UNESCO	Unión Europea
Creación	1977	1984
Objeto	Análisis temático. Búsqueda de documentos	Tratamiento de información generada internamente
Campos	Multidisplinar	Multidisplinar
Jerarquía	Monojerárquico	Polijerárquico
Idiomas	Español, inglés, francés y ruso	22 lenguas UE + Croata + Serbio
Actualización	2008	2012

Tabla 1. *Características de Tunesco y Teurovoc.*

Ahora bien, un tesauro sólo es útil si los conceptos, términos (descriptores y no descriptores) y relaciones que recoge son revisados con la frecuencia necesaria, incorporando "la terminología derivada del desarrollo de la ciencia o materia a la que se dedica (…), [cubriendo] lagunas o fallos detectados durante su utilización [y adaptando] a las necesidades de recuperación manifestadas por los usuarios" (Pérez Agüera, 2004). Lo contrario da lugar a vocabularios desactualizados, incapaces de cumplir adecuadamente con su función, ya que carecen de los términos necesarios para representar el contenido de los documentos. Pérez Agüera (2004) señala esta circunstancia como un problema habitual de estas herramientas, mientras que Fernández-Quijada (2012) manifiesta, por su parte, la falta de una actualización de Tunesco "que responda a la naturaleza cambiante de los ámbitos sociales y de los sectores profesionales".

Este trabajo tiene como objetivos analizar el grado de actualización de los macro-tesauros (1) indicados, así como determinar la capacidad de las noticias de divulgación científica de ser utilizadas para esa tarea de renovación, a partir de su indización. Como consecuencia de los anteriores, se plantea sí estos vocabularios resultan adecuados para describir noticias de divulgación científico-tecnológica.

El presente artículo se organiza de la siguiente forma: La metodología se centra en la descripción de las características de las noticias de divulgación científico-tecnológica así como en el proceso de indización de titulares de periódicos y sistema de clasificación de resultados. Posteriormente se detalla el funcionamiento de los dos métodos empleados, N-gramas y Apache Solr, para la búsqueda de equivalencias entre términos extraídos y términos presentes en los vocabularios Tunesco y Teurovoc y la clasificación de términos e función de los datos obtenidos. A continuación, se presentan los resultados. En la discusión se analizan algunos datos y sus implicaciones. Finalmente, se establecen una serie de conclusiones, en relación con los objetivos definidos y futuras líneas de investigación.

2. Metodología.

Una de las formas de evaluación de la calidad de un tesauro es la evaluación extrínseca, definida por Gil Leiva (2008) y centrada en el estudio del comportamiento de dicho vocabulario en la indización y recuperación de información.

Desde este punto de vista, podemos considerar el grado de actualización como un indicador de calidad extrínseca, considerando como tal la capacidad de un tesauro para describir adecuadamente un documento reciente.

Conocida la amplia utilización de los tesauros Tunesco y Teurovoc fuera de sus ámbitos de actuación (las instituciones en que se circunscriben), nos planteamos recurrir a documentos externos al sistema para medir el grado de actualización antes indicado. En este caso, acudimos a noticias de divulgación científico-tecnológico publicadas en diarios internacionales, ya que pueden constituir una fuente de renovación adecuada teniendo en cuenta las características de la información periodística también definidas por (García Gutiérrez y Lucas Fernández, 1987, p. 19-27); (Cebrián, 1997) y (Rubio Lacoba, 2007), entre muchos otros:

1) Corpus de menor extensión que, en general, cualquier documento interno de la UNESCO o Unión Europea;
2) Actualidad e inmediatez;
3) Enciclopedismo: cualquier término es susceptible de interés para el medio y su público;
4) Universalidad de la procedencia de la noticia y de la fuente, de los canales de transmisión y de las audiencias;
5) Proximidad: un asunto de carácter noticioso

es cubierto por medios internacionales, a cuya información se puede acceder vía web en cualquier momento y lugar;

6) Contraste de la información, que constituye uno de los pilares de la labor periodística;

7) Normalización, mediante la guía de estilo propia de cada medio;

Centrándonos en las noticias especializadas en ciencia y tecnología:

8) Interés divulgativo y pedagógico;

9) La utilización de fuentes acreditadas, ya que las noticias de divulgación habitualmente parten de información publicada en revistas científicas;

10) Uso y adaptación de lenguajes técnicos (Castillo Blasco, 2006) para "recontextualizar aspectos del conocimiento o de la práctica científica" (Alcíbar Cuello, 2004).

Definido el objeto de análisis, se diseñó una muestra a partir de la selección sucesiva de países, periódicos y titulares. A partir de cinco parámetros previamente definidos (2), se estableció un ranking de países punteros en Ciencia y Tecnología, seleccionando aquellos situados en los primeros puestos de, al menos, tres parámetros, a saber: Gasto Interior Bruto en I+D+i ($); valor bruto ($) de exportaciones de alta tecnología; valor bruto recibido por cada país en concepto de royalties, cánones y otros por uso de patentes ($); número de artículos publicados por país; total de premios recibidos más relevantes en Ciencia y Tecnología por país (3). Finalmente, los países escogidos fueron Alemania, Canadá, China, República de Corea, España, Estados Unidos, Francia, Italia, Japón, Reino Unido y Rusia. Este último país, aunque únicamente cumplía uno de los requisitos, se ha incluido en el estudio debido a su relevancia en el ámbito científico, político y económico.

Posteriormente, a partir de los datos de la web de rankings *4International Media & Newspaper* (4), se seleccionó el periódico de información general más popular en cada país, teniendo en cuenta datos de lectura y acceso tanto de las ediciones digitales como impresa. En el caso de Estados Unidos y China se seleccionaron dos medios (5). Después se seleccionaron las noticias más destacadas de las secciones de Ciencia y/o Tecnología de cada medio, una por día y sección, durante cuatro meses, comprendidos entre el 9 de marzo y 9 de julio de 2012. Así, se obtuvo un corpus de 1599 noticias del que se seleccionaron aleatoriamente dos sub-muestras: la sub-muestra 1 (M1), se compone de 159 artículos (10%), la segunda (M2), está formada por 320 titulares (20%). La distribución de noticias en cada muestra se muestra en la siguiente tabla.

País	*Medio*	*M1*	*M2*
Alemania	*Süddeutsche Zeitung*	*8*	*19*
Canadá	*The Global and Mail*	*11*	*21*
China	*China Daily*	*7*	*13*
	The China Post	*13*	*21*
Corea	*The Korea Times*	*10*	*22*
España	*El Mundo*	*12*	*19*
EE.UU	*The New York Times*	*9*	*41*
	The Washington Post	*21*	*38*
Francia	*Le Monde*	*18*	*41*
Italia	*La Reppublica*	*21*	*5*
Japón	*Yomiuri Shimbun*	*0*	*40*
Reino Unido	*The Daily Telegraph*	*23*	*11*
Rusia	*Pravda*	*6*	*29*

Tabla 2. *Número de noticias por diario, seleccionadas aleatoriamente*

A continuación, se indizaron las noticias de cada conjunto, manualmente y en lenguaje natural, extrayendo entre 1 y 6 palabras clave de cada titular (6) (7) y se tradujeron a Español, Francés e Inglés, empleando herramientas de uso común, como *Word Reference*, *Linguee o Google Translator*. Con ello disponemos de un mecanismo que permite el uso cruzado de dichos idiomas como lenguajes pivote para desambiguar casos de homonimia y polisemia (Areas da Luz Fontes et al., 2010; Degani & Tokowicz, 2010; Marchisio & Liang, 2001).

Español	*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*

Tabla 3. Muestra de términos extraídos en español, inglés y francés

Más adelante, se llevó a cabo una búsqueda de equivalencias de las palabras clave extraídas con respecto a cada macro-tesauro. Para ello se construyeron sendas colecciones de documentos a partir de la terminología de cada vocabulario y se indizaron utilizando dos técnicas caracterizadas por su simplicidad y amplia utilización: N-gramas y el sistema de recuperación de información Apache Solr. De esta forma, en un único documento y para cada colección, se agruparon tanto descriptores como no-descriptores de cada concepto en español, inglés y francés (en campos separados). Posteriormente se buscaron automáticamente las equivalencias entre términos extraídos y términos de los tesauros. La siguiente figura muestra

la relación entre palabra clave extraída y términos en tesauro para el cálculo del *score* de similitud.

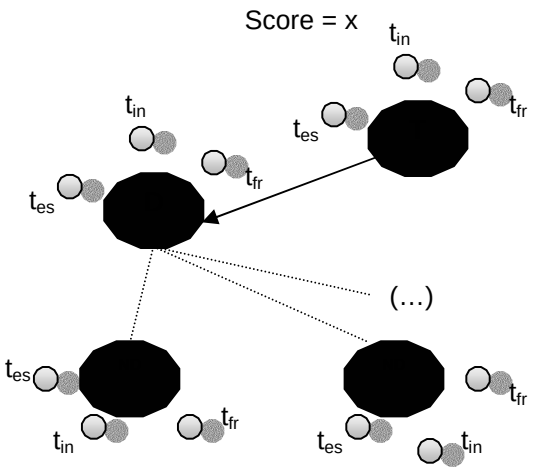

Figura 1. Relación término extraído y término de macro-tesauro (descriptores y no descriptores). Elaboración propia a partir de Smits & Friis-Christensen (2007)

Para evaluar los resultados obtenidos y determinar el grado de equivalencia entre términos, clasificados de la siguiente forma:

- Término correcto (TC): Los términos de los titulares y del tesauro coinciden tanto en significante como en significado.
- Término sinónimo (TS): Los significantes son diferentes pero hacen referencia al mismo significado.
- Término específico (TE): El término del titular es más específico que el del término en el tesauro.
- Término genérico (TG): El significante del titular es más genérico que el del tesauro (8).
- Término relacionado (TR): Los significados mantienen cierta relación semántica.
- Término nuevo (TN): No halla relación entre con los términos del tesauro.
- Término con falsa equivalencia (TFE): El término del titular no mantiene relación semántica con el devuelto, constituyendo, por tanto, una relación de falsa equivalencia que, en última instancia, se traduce en TN.

Finalmente, se calcularon estadísticas para cada tipo de relación entre pares de términos, así como la precisión exacta, cercana y total.

3. El método N-gramas.

Esta técnica emplea el modelo vectorial basado en el cálculo de la semejanza entre vectores, tal como lo aplican Aronson et al. (2000), Pouliquen et al. (2002), Tolosa y Bordignon (2008) y Rodríguez-Torrejón & Martín-Ramos (2012). Se ha utilizado la sub-muestra M1, compuesta de 301 palabras clave únicas, extraídas de 159 noticias (10% del total de 1599 obtenidas).

La distribución de noticias por periódico se detalla en la Tabla 2. Se obtuvieron 508 palabras clave, reducidas a 301 al eliminar duplicados. No se aplicaron procesos de *stemming* o lematización. Primero se elaboraron 3 conjuntos de documentos de palabras clave extraídas y los términos de Tunesco y Teurovoc (descriptores y no descriptores), en los tres idiomas antes indicados. Posteriormente se comparó cada palabra clave extraída con cada término de cada vocabulario, siguiendo este proceso:

1) Identificación de los bi-gramas únicos solapados de los términos a comparar;
2) Aplicación de la fórmula del Coeficiente de similitud de Dice (C_d) a cada palabra clave extraída y expresiones que integran cada macro-tesauro, en el mismo idioma. El resultado, para cada término, oscila entre 0 (ninguna semejanza) y 1 (similitud exacta);
3) Cálculo de la media (M_{si}) de los C_d para cada palabra clave. El término de Tunesco o Teurovoc, en cada caso, seleccionado por el script será aquel con la media (M_{si}) de similitud más elevada;
4) Clasificación de las medias (M_{si}) en intervalos agrupados de similitud: [0 – 0,6), [0,6 – 0,8) y [0,8 – 1]. En la Tabla 4 se presenta la distribución de los términos por intervalos. La descripción detallada de este proceso y ejemplo se incluye en el Apéndice.

Intervalos (9)	*Tunesco*	*Teurovoc*	*Intervalos agrupados*	
			UN	*EU*
[0,0-0,4)	14	9	91	57
[0,4-0,5)	29	19		
[0,5-0,6)	48	29		
[0,6-0,7)	34	47	59	80
[0,7-0,8)	26	33		
[0,8-0,9)	18	17	151	164
[0,9-1]	132	147		

Tabla 4. Intervalos de semejanza de palabras claves extraídas y términos Tunesco y Teurovoc

Los resultados obtenidos (Tabla 5) con este método muestran que Teurovoc está, en general, más actualizado que Tunesco, ya que en los intervalos más altos (que agrupan los términos con mayor similitud), recoge más elementos.

	Tunesco		*Teurovoc*	
	Total	*Total*	*Total*	*%*
TC	161	53,49	176	58,47
TS	7	2,33	4	1,33
TG	7	2,33	9	2,99
TE	15	4,98	11	3,65
TR	28	9,30	30	9,97
TFE	83	27,57	69	22,92

Tabla 5. Equivalencias obtenidas por N-gramas para los términos de Tunesco y Teurovoc

Por otro lado, el método n-gramas identifica correctamente más de un 50% de términos extraídos en cada macro-tesauro (TC). El resto de palabras clave (menos de un 20%) guardan otros tipos de relaciones. Finalmente, en ambos vocabularios, más de un 20% de similitudes (TFE) fueron mal atribuidas porque buena parte de los resultados se agrupaban en los intervalos más bajos, [0,0 y 0,6), donde la similitud entre pares de términos es menor.

4. El método de Apache Solr.

Con los datos obtenidos empleando la técnica anterior, se planteó la necesidad de profundizar sobre los resultados TFE, por lo que se procedió a realizar una búsqueda de equivalencias entre pares de términos utilizado otro método diferente. En este caso se partió de la segunda submuestra M2, compuesta de 320 titulares (20%) y de un total de 1014 palabras clave extraídas, que se redujeron a 599 términos únicos al suprimir duplicados.

La distribución de noticias por periódico se detalla en la Tabla 2. Para ello se construyeron sendas colecciones de documentos a partir de la terminología de Tunesco y Teurovoc, donde los términos de cada uno se identificaron como documentos para su indización mediante Apache Solr (10), siguiendo la estructura indicada en esta Tabla:

Campo	*Descripción*
id	Identificador del concepto
type	Tesauro (Tunesco ó Teurovoc)
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 6. Estructura de campos para la indización de términos como documentos en Apache Solr

Después se efectuaron una serie de búsquedas automáticas para cada término extraído e idioma en cada colección del macro-tesauro. Tras diversos ensayos se configuró un procedimiento compuesto de 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).

De estas consultas, Apache Solr proporciona un medida de similitud o *score* (11) de la consulta con cada uno de los documentos (los términos) del sistema. Es evidente que las equivalencias literales exactas de los términos de los titulares con descriptores y no descriptores permiten determinar una identificación exacta o muy cercana. Por este motivo, a los *score* 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. De forma experimental también se comprobó la necesidad de potenciar los resultados de la consulta Q1, por lo que se aplicó un *boost* de 2,5. Por otro lado, para las búsquedas Q1, Q4, Q5, Q6 y Q7, a la vista de los datos de ensayos previos, se estableció para un umbral mínimo de score por debajo del cual debían desecharse dichos resultados. Finalmente, los resultados obtenidos (Tabla 7) se analizaron para determinar el tipo de equivalencia entre los términos de los titulares y los recogidos en los tesauros y se clasificaron de la siguiente forma:

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

Tabla 7. Equivalencias obtenidas por Apache Solr para los términos de Tunesco y Teurovoc

Por lo que respecta a Tunesco: algo más del 44% de los términos tenían una equivalencia exacta (TC), mientras que casi un 16% de las mantienen una relación de jerarquía (TG y TE) y poco más de un 12% guardan una relación de tipo asociativo (TR). Además, Apache Solr identificó más de un 25% de términos para los que no halló relación (TN) o era falsa (TFE). Respecto a las equivalencias con Teurovoc: para casi la mitad de los términos (48,41%) se encontró una equivalencia exacta (TC), más de un 14% guardan una relación de jerarquía (TG y TE) y casi un 12%, es de tipo asociativo (TR). Para algo más de un 24% Apache Solr no halló equivalente (TN) o era falsa equivalencia (TFE). Se puede afirmar que siempre es preferible una equivalencia exacta (TC) a una de sinonimia (TS). También que es mejor una equivalencia de TE a una TG, ya que en el primer caso, el significado del término de un titular es cubierto por uno del tesauro, que no sucede al contrario.

5. Resultados.

A partir de los datos anteriores y para evaluar la eficacia de los procedimientos utilizados, se calculó la precisión (*P*) de las búsquedas efectuadas en cada macro-tesauro (Cleverdon & Keen, 1966; Lancaster, 2002; Owen & Cochrane, 2004; Tolosa & Bordignon, 2008; y Hage et al., 2010). Considerando TR_{REL} y TR_{TOT} como términos recuperados relevantes y el total de términos recuperados, respectivamente, se tendría la siguiente ecuación:

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

Se realizaron varios cálculos de precisión: precisión exacta (P_{EX}), que consideraría como relevantes únicamente los términos correctos (TC); precisión cercana (P_{CLOSE}), que también tendría en cuenta los sinónimos (TC + TS); y la precisión total (P_{TOTAL}), que incluiría cualquier tipo de relación (TC + TS + TG + TE + TR).

De este modo, se obtendrían los valores de precisión señalados en las Tablas 8 y 9. En el caso de Tunesco (Tabla 8), la precisión exacta (P_{EX}) y la cercana (P_{CLOSE}) son relativamente bajas, aunque la técnica n-gramas consigue mejores resultados. La precisión total (P_{TOTAL}) es similar en ambos casos y no llega a 0,75. Alrededor de un 25% de términos no han sido recuperados por ninguno de los métodos.

	N-gramas	*Apache Solr*
P_{EX}	0,5348	0,4407
P_{CLOSE}	0,5581	0,4641
P_{TOT}	0,7242	0,7462

Tabla 8. Precisión exacta, cercana y total obtenidas con N-gramas y Apache Solr para Tunesco

La representación gráfica de estos datos se muestra en el Gráfico 1:

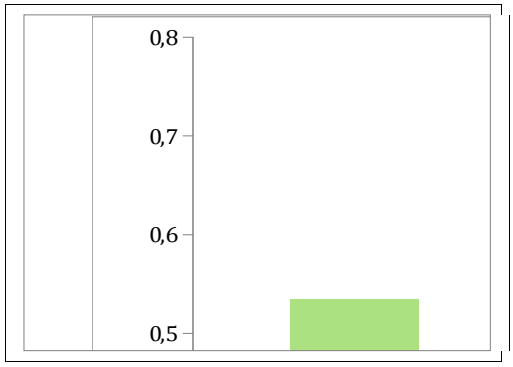

Figura 2. Resultados de precisión para Tunesco

En cuanto a Teurovoc, la precisión exacta (P_{EX}) y la cercana (P_{CLOSE}) tampoco son elevadas, aunque sí algo mejores que las obtenidas para Tunesco. En este caso también se consiguen mejores resultados con la técnica n-gramas. Por otro lado, la precisión total (P_{TOTAL}) es prácticamente igual para las dos técnicas y sobrepasan ligeramente el 0,75. Poco menos de un 25% de los términos no han sido recuperados por ninguno de los métodos.

	N-gramas	*Apache Solr*
P_{EX}	0,5847	0,4841
P_{CLOSE}	0,5980	0,5008
P_{TOT}	0,7641	0,7595

Tabla 9. Precisión exacta, cercana y total obtenidas con N-gramas y Apache Solr para Teurovoc

Una representación gráfica de estos datos se muestra en el Gráfico 2:

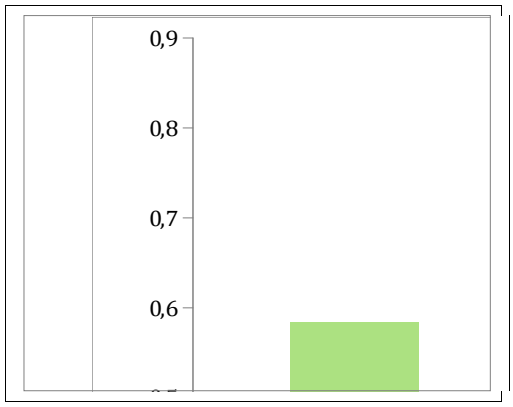

Figura 3. Resultados de precisión para Teurovoc

Habitualmente el cálculo de la precisión (*P*), va ligado al de la exhaustividad (*E*), de forma que es posible determinar la Medida F o Armónica (Van Rijsbergen, 1979, p. 129-135). En este trabajo no se calcula la exhaustividad de la recuperación de información, ya que los datos negativos (clasificados como TN y TFE) han sido ana-

lizados individualmente para ambas técnicas y no se han hallado términos de equivalencia exacta (TC). Parece evidente entonces que la exhaustividad tendería a 1. Por otro lado, se ha comprobado que, en la mayoría de casos, existían términos genéricos (TG) que se podrían utilizar para describir las noticias, en lugar de los TN y TFE detectados.

6. Discusión.

La especialización de los campos elegidos para este estudio puede haber influido en los resultados. Un simple vistazo permite comprobar que Tunesco recoge más términos y expresiones de ciencia y educación, mientras que Teurovoc (probablemente derivado del desarrollo legislativo en materia de administración electrónica, protección de datos, descarga de información banca electrónica, etc.) se ha especializado más en el ámbito tecnológico.

Un ejemplo claro lo encontramos en el término macho alfa, que podría ser recogido en Tunesco, por estar más enfocado a Ciencia, pero no en Teurovoc. También podría ocurrir que las noticias de divulgación cada vez más se circunscriban al ámbito tecnológico, lo que explicaría porqué este último vocabulario obtiene mejores resultados de precisión.

Por otro lado, la utilización de noticias especializadas en otras áreas podría producir tasas de precisión diferentes. Así, el uso de artículos de las secciones de nacional o internacional, casi siempre de corte político y/o económico, probablemente habría incrementado los resultados para Teurovoc

La existencia de TN y TFE constituye una verdadera fuente para la actualización de los macro-tesauros, ya que son términos no recogidos en los mismos. Igual ocurre con los TG, TE y TR, pues, aunque ambas técnicas detectan relaciones jerárquicas y asociativas, los términos de Tunesco y Teurovoc propuestos no son los más adecuados para la indización de estas noticias.

Alrededor de 3/4 de las palabras clave extraídas guardan algún tipo de relación (exacta, sinonimia, jerárquica o asociativa) tanto con los términos de Teurovoc como con los de Tunesco. Ahora bien, es cierto que ambas técnicas, basadas en el cálculo de la similitud morfológica entre pares de términos, no establecen relaciones semánticas en ausencia de esta semejanza. Esto justificaría, al menos en parte, la falta de equivalencias para los TN (términos nuevos) determinados por Apache Solr, así como los falsos equivalentes detectados (TFE) por ambos métodos.

En este sentido, sería conveniente analizar individualmente cada palabra clave en el contexto de Tunesco y Teurovoc y comprobar si existen otras relaciones semánticas, así como incluir términos nuevos en los tesauros si es necesario.

Aunque no se ha profundizado en las relaciones entre los términos, descriptores y no descriptores, presentes en los tesauros, se aprecia la necesidad de revisar éstas.

La consideración de no descriptor genera ruido cuando se intentan recuperar documentos a partir de ciertos términos que actualmente son utilizados abundantemente y que, por tanto, deberían ser elevados a la categoría de descriptor. Tal podría ser el caso *smartphone* (no descriptor) y teléfono móvil (descriptor).

Por otro lado, no hay que olvidar la subjetividad subyacente en todo proceso de indización humana, especialmente cuando se realiza en lenguaje natural y no se está sujeto a las restricciones propias de un lenguaje controlado como un tesauro. Por esta razón es importante tener en cuenta los factores que, según Nakurawa (2009), influyen en la indización (Tabla 10):

Factor	*Descripción*
Qué se indiza	Titulares de noticias. Si no eran descriptivos se descartaban, salvo que hubieran sido cubiertas previamente (y por tanto, conocida)
Quién indiza	Formación, experiencia, conocimiento del asunto y dominio de la herramienta de indización son factores clave. Ya se había trabajado antes con ambos macro-tesauros
Política de indización	Como norma general, siguiendo a Currás (1991), se utiliza, en general la forma singular. En cuanto a la carga de trabajo, se decidió dedicar de 2 a 3 minutos por noticia

Tabla 10. Factores que influyen en la indización humana

7. Conclusiones.

A la vista de los datos de precisión obtenidos, ningún macro-tesauro permite describir correctamente noticias de divulgación científico-tecnológica, por la ausencia de los términos más adecuados para una indización pertinente. El empleo de otros términos con los que las palabras clave extraídas guardan una relación diferente a la exacta o de sinonimia provocaría ruido, al aportar resultados poco relevantes respecto de la información que se busca o silencio, al no recuperar datos.

Sin embargo, como ha ocurrido en ocasiones anteriores, sí que pueden utilizarse como base para la construcción de otras herramientas de organización del conocimiento, previa adapta-

ción a las necesidades de sus usuarios y empleando adicionalmente otros vocabularios y fuentes para conseguir completitud adecuada. Con respecto al grado de actualización de Tunesco y Teurovoc, la precisión exacta (P_{EX}) y la cercana (P_{CLOSE}) no son elevadas en ningún caso, aunque los datos son mejores en el caso de Teurovoc. Una revisión más frecuente de ambas herramientas es necesaria y obligatoria, de forma que se cuestionen, entre otros, analicen e incrementen las tasas de equivalencia (relación entre número de no descriptores y descriptores) y enriquecimiento (proporción entre número de relaciones jerárquicas y asociativas, y la cifra total de descriptores), definidas por Slype (1991, en García Jiménez, 2002, pp. 121-122).

La inclusión de nuevos términos, y la redefinición de las relaciones entre los existentes incrementarían indudablemente la calidad de las herramientas analizadas. Como señala Currás (2010), "la evolución y el progreso se mueven a gran velocidad en todas las áreas, y por eso parece que, cuando un concepto, idea o forma de pensar se ha establecido y consolidado, nos cruzamos con algo nuevo que, al menos, atrae nuestra atención e incluso contrasta totalmente con lo previo".

Por otro lado, las noticias de divulgación científico-tecnológica constituyen una fuente adecuada para la actualización de estos macro-tesauros, ya que parte de la evolución de la ciencia y tecnología se refleja en los diarios internacionales. De esta forma, este tipo de información se puede emplear para la inclusión de nuevos conceptos y/o términos, así como para la redefinición de las relaciones entre éstos. Además, la especialización de los artículos utilizados facilita la renovación de micro-tesauros concretos. Noticias de prensa publicadas en otras secciones (como economía o cultura) podrían emplearse para actualizar áreas específicas de tesauros y/o micro-tesauros no contempladas en este trabajo. Asimismo, podría emplearse la información periodística como fuente para desarrollar y/o actualizar herramientas de organización del conocimiento distintas a los tesauros, como, por ejemplo, las ontologías, donde las relaciones entre conceptos son mucho más complejas y formales.

En ambos vocabularios, los lenguajes pivote se utilizaron para resolver los asuntos de homonimia (12) y polisemia. Ahora bien, en el caso del método n-gramas, el efecto de la paronimia (expresiones muy parecidas de las que sólo cambia alguna letra) ha dado lugar a términos falsos equivalentes (TFE) para los que podría haberse encontrado una relación jerárquica (TG, TE) o asociativa (TR) apropiada. Por ejemplo, corrupción política (palabra clave extraída) y coalición política (término Teurovoc). El término corrupción (que sí ha obtenido Apache Solr) es más adecuado y guarda una relación de término genérico (TG). El aprovechamiento de las propias relaciones semánticas internas entre conceptos de los tesauros también podría ser útil cuando la desambiguación es necesaria. Profundizar en esta área constituye una línea futura para mejorar el método.

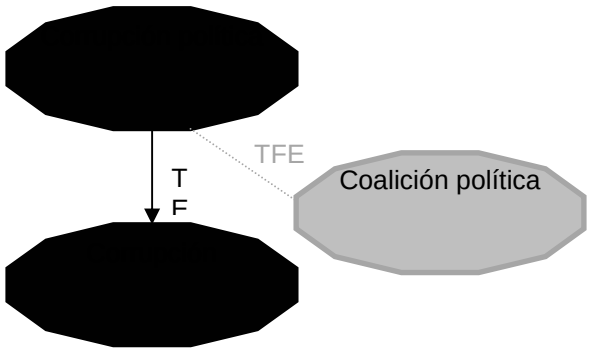

Figura 4. Relación término extraído y términos en Teurovoc con método N-gramas

Además, los dos vocabularios incluyen un volumen diferente de descriptores y no descriptores, dependiendo de la riqueza léxica (sinónimos y cuasi-sinónimos) del idioma.

Los lenguajes pivote también han sito útiles en estos casos, devolviendo el término correcto aún cuando no se recogía en alguno de los idiomas analizados. Así ocurre con *wildlife* (palabra clave extraída) y *wild animal* (término en Teurovoc).

La lematización se revela como paso previo necesario a la aplicación de cualquier técnica. En el método n-gramas no se ha utilizado y ha provocado que algunos términos se ubiquen en intervalos bajos cuando en realidad son términos correctos (TC) y forman parte de los intervalos más altos. Esto habría reducido el proceso de revisión posterior. Un ejemplo lo encontramos en los términos *city* y *cities* o *lobby* y *lobbies*.

Aunque la utilización de lenguajes pivote ha resuelto buena parte de esta problemática, un estudio posterior podría trabajar en esta cuestión.Finalmente, los resultados obtenidos mediante N-gramas y Apache Solr, muestran una diferencia de ± 10, tanto para la precisión exacta (P_{EX}) como la cercana (P_{CLOSE}), mientras que la total (P_{TOTAL}) es muy similar. Estudiar los resultados concretos para cada par de términos, es necesario para profundizar en la disparidad de resultados afinar ambos métodos.

Notas.

(1) Tesauro que incluye varios micro-tesauros de distintos campos temáticos.

(2) Los datos de los cuatro primeros parámetros proceden de Banco Mundial (2012), año 2008.

(3) 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.

(4) Disponible en: http://www.4imn.com. Esta web realiza un ranking *basado en la popularidad de las webs de los diarios* a través de *un algoritmo que incluye datos,* a partir de tres motores de búsqueda: Google Page Rank, Majestic Seo Referring Domains y Alexa Traffic Rank. Desde 4IMN podemos conocer los periódicos más populares por país o de acuerdo a seis divisiones territoriales (Norteamérica, Sudamérica, Europa, Oceanía, Asia, África).

(5) Estados Unidos es el país más destacado en todos los parámetros; el segundo es el país con previsiones de PIB más elevado en 2014, según IMF (2013), lo que puede implicar una importante inversión en Ciencia y Tecnología en el futuro.

(6) Algunos de los medios seleccionados limitan el acceso a noticias a texto completo, debido a su modelo de negocio. Sin embargo, sí podemos conocer sus titulares, que constituyen una de las partes esenciales del artículo y que resume en pocas palabras su contenido. Por esta razón se establece el titular como unidad mínima de análisis.

(7) Loza Mencía & Füernkranz, 2008, asignan 5,37 en su estudio.

(8) 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.

(9) El grado de similitud mínima entre términos es de 0,4, por lo que el primer intervalo es [0,0-0,4).

(10) Apache Solr es una plataforma de código abierto para el desarrollo de motores de búsqueda web que está basado en el software de recuperación de información Apache Lucene. Más información en: http://lucene.apache.org/Solr/

(11) 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. También puede consultarse Loza Mencía & Füernkranz (2008), Campos & Romero (2008), Shvaiko et al. (2010), entre otros.

(12) Este fenómeno ocurre entre dos términos que se escriben (homógrafos) o suenan (homófonos) igual, pero tienen etimología diferente.

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. *Revista Digital do PPGL* 3 (1). Recuperado el 07/06/2013 de http://revistaseletronicas.pucrs.br/ojs/index.php/letronica/article/view/7074

Aronson, A. R., Bodenreider, O., Chang, H. F., Humphrey, S. M., Mork, J. G., Nelson, S. J. and Wilbur, W. J. (2000). The NLM Indexing Initiative. *Proceedings of AMIA Annual Symposium*, 17–21.

Banco Mundial. (2012). *Indicadores del desarrollo mundial: PIB (US$ a precios actuales) (Estadística)*. Recuperado el 01/05/2013 de http://datos.bancomundial.org/indicador/NY.GDP.MKTP.CD

Baxter, R., Blomeley, F. & Kemsley, R. (2002). *The AIM25 Project. Ariadne*, Issue 31. Recuperado el 01/04/2013 de http://www.ariadne.ac.uk/issue31/aim25/

Campos, L. M. de & Romero, A. E. (2009). Bayesian network models for hierarchical text classification from a thesaurus. *Special Section on Graphical Models and Information Retrieval*, 50 (7), 932–944. doi:10.1016/j.ijar.2008.10.006

Carrizo Sainero, G. (2000). *La información en ciencias sociales*. Gijón: Trea.

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 el 04/04/2013 de http://www.tdx.cat/bitstream/handle/10803/9982/castillo.pdf?sequence=1

Cebrián, B. J. (1997). *Fuentes de consulta para la documentación informativa*. Madrid: Universidad Europea - CEES.

Cleverdon, C. W., & Keen, M. (1966). *Aslib Cranfield research project - Factors determining the performance of indexing systems; Volume 2, Test results*. Crandfield: National Science Foundation. Recuperado 04/04/2013 de http://dspace.lib.cranfield.ac.uk/handle/1826/863

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

Currás, E. (2010). *Ontologies, taxonomies and thesauri in systems science and systematics*. Oxford: Chandos Publishing.

Daudaravicius, V. (2010). The influence of collocation segmentation and top 10 items to keyword assignment performance. In *Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing* (648–660). Berlin, Heidelberg: Springer-Verlag. Recuperado 11/04/2013 de http://link.springer.com/chapter/10.1007/978-3-642-12116-6_55

Degani, T., & Tokowicz, N. (2010). *Semantic ambiguity within and across languages: an integrative review*. Quarterly Journal of Experimental Psychol-

ogy 63 (7), 1266–1303. Recuperado 04/04/203 de http://www.ncbi.nlm.nih.gov/pubmed/19953429

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 el 06/06/2013 de http://www.ub.edu/bid/29/fernandez2.htm

Fiser, D., & Sagot, B. (2008). Combining Multiple Resources to Build Reliable Wordnets. In Sojka, A. Horak, I. Kopecek, & K. Pala (Eds.), *Text, Speech and Dialogue, Proceedings* (Vol. 5246, 61–68). Berlin: Springer-Verlag Berlin. Recuperado el 07/06/2013 de http://link.springer.com/chapter/10.1007%2F978-3-540-87391-4_10#page-1

García Gutiérrez, A., & Lucas Fernández, R. (1987). *Documentación Automatizada en los Medios Informativos*. Madrid: Paraninfo.

García Jiménez, A. (2002). *Organización y gestión del conocimiento en la comunicación*. Gijón: Trea.

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. Recuperado de http://www.tandfonline.com/doi/abs/10.1080/00379810050006902#.UdgemD7AWmc

Gil Leiva, I. (2008). *Manual de indización: teoría y práctica*. Gijón: Trea.

Hage, W. R. van, 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. Recuperado el 06/06/2013 de http://www.cs.vu.nl/~guus/papers/Hage10d.pdf

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. Recuperado el 11/06/2013 de http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=1491146&contentType=Conference+Publications

Lancaster, F. W. (2002). *El control del vocabulario en la recuperación de información* (2a ed.). Valencia: Universidad de Valencia.

Loza Mencía, E., & Füernkranz, J. (2008). Efficient Pairwise Multilabel Classification for Large-Scale Problems in the Legal Domain. In W. Daelemans, B. Goethals, & K. Morik (Eds.), *Machine Learning and Knowledge Discovery in Databases, Part II, Proceedings* (Vol. 5212, 50–65). Berlin: Springer-Verlag Berlin. Recuperado el 04/04/2013 de http://www.ke.tu-darmstadt.de/~juffi/publications/ecml-pkdd-08.pdf

Marchisio, G. B., & Liang, jisheng. (2001). Experiments in Trilingual Cross-Language Information Retrieval. In *Proceedings 2001 Symposium on Document Image Understanding Technology* (169–179). Universty of Maryland. Recuperado el 04/04/2013 de http://bitly.es/nr

Martínez, A. M., Ristuccia, C. A., Stubbs, E. A., Valdez, J. C., Gamba, V. L., Mendes, P. V., Caminotti, M. L. (2011). La estructura sistemática del tesauro: indicadores para evaluar su calidad. *Revista Española de Documentación Científica,* 34 (1), 29–43. doi:10.3989/redc.2011.1.765. Recuperado 05/04/2013 de http://redc.revistas.csic.es/index.php/redc/article/viewArticle/681

Mastora, A., Monopoli, M., & Kapidakis, S. (2008). *Term Selection Patterns for Formulating Queries: a User Study Focused on Term Semantics*. New York: IEEE. Recuperado el 04/04/2013 de http://www.ionio.gr/~sarantos/repository/c45C-ICDIM2008MasMon.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 el 04/04/2013 de http://www.ies.ufpb.br/ojs/index.php/ies/article/view/2925

Orenga-Gaya, L., & Giralt, O. (2011). The official gazette of the Generalitat de Catalunya: genesis of a digital newspaper. El Profesional de la Información, 20 (3), 340–344. Recuperado el 01/04/2013 de http://www.doc6.es/media/pdfs/articulos/diario_digital.pdf

Owens, L. A., & Cochrane, P. A. (2004). Thesaurus evaluation: Review, reinassance y revision. In *The Thesaurus: Review, Renaissance, and Revision*. Routledge. Recuperado el 04/04/2013 de http://bitly.es/nt

Pérez Agüera, J. R. (2004). Automatización de tesauros y su utilización en la web semántica. *BiD: textos universitaris de biblioteconomia i documentació* 13. Recuperado el 08/03/2013 de http://www.ub.edu/bid/13perez2.htm

Pollit, A. S., Ellis, G. P., & Smith, M. P. (1995). Using the thesaurus to view and filter environmental databases - an example using Eurovoc to search epoque - the European Parliament Online Query System. In P. Stancikova & I. Dahlberg (Eds.), *Environmental Knowledge Organization and Information Management, Supplement Vol* 1 (21–32). Frankfurt: Indeks Verlag.

Pouliquen, B., Steinberger, R., & Ignat, C. (2004). Automatic linking of similar texts across languages. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing III* Vol. 260 (307–316). Amsterdam Me: John Benjamins B V Publ.

Pouliquen, B., Delamarre, D., & Le Beux, P. (2002). Indexation de textes médicaux par extraction de concepts, et ses utilisations. In A. Morin & P. Sébillot (Eds.), *6eme Conférence Internationale « Journées d'Analyse de Données textuelles »* Vol. 2 (17–628). Presentado en the JADT'2002, St. Malo, France. Recuperado 09/05/2013 de http://www.med.univ-rennes1.fr/~poulique/nomindex.pdf

Rodríguez Mateos, D. (2011). Internet y su influencia sobre la documentación audiovisual. In *Documentación audiovisual: Nuevas tendencias en el entorno digital.* Madrid: Síntesis.

Rodríguez-Torrejón, D. A., & Martín-Ramos, J. M. (2012). N-gramas de Contexto Cercano para me-

jorar la Detección de Plagio. In *Actas del II Congreso Español de Recuperación de Información (CERI-2012)*. Valencia: Universitat Jaume I. Recuperado el 04/07/2013 de http://users.dsic.upv.es/grupos/nle/ceri/papers/ceri2012_torrejon_ramos_ngrams.pdf

Rubio Lacoba, M. (2007). *Documentación informativa en el periodismo digital*. Madrid: Síntesis. Recuperado el 05/06/2013 de http://www.marcialpons.es/libros/documentacion-informativa-en-el-periodismo-digital/9788497564595/

Saric, F., Snajder, J., Basic, B. D., & Eklic, H. (2005). Enhanced thesaurus terms extraction for document indexing. In Proceedings of the 27th International Conference on Information Technology Interfaces (214 - 219) . Recuperado el 02/06/2013 de http://www.med.univ-rennes1.fr/~poulique/nomindex.pdf

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. Recuperado el 04/06/2013 de http://www.emeraldinsight.com/journals.htm?articleid=862260

Shvaiko, P., Oltramari, A., Cuel, R., Pozza, D., & Angelini, G. (2010). Generating Innovation with Semantically Enabled TasLab Portal. In L. Aroyo, G. Antoniou, E. Hyvonen, A. TenTeije, H. Stuckenschmidt, L. Crabral, & T. Tudorache (Eds.), *Semantic Web: Research and Applications, Pt 1, Proceedings* Vol. 6088 (348–363). Recuperado el 06-06-2013 de http://www.loa.istc.cnr.it/Papers/TasLabPortal_final.pdf

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

Smits, P. C., & Friis-Christensen, A. (2007). Resource discovery in a European Spatial Data Infrastructure. *IEEE Transactions on Knowledge and Data Engineering* 19 (1), 85–95. doi:10.1109/TKDE.2007.250587

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 el 03/06/2013 de http://hdl.handle.net/10760/12243

Van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Newton: Butterworth-Heinemann. Recuperado el 01/06/2013 de http://openlib.org/home/krichel/courses/lis618/readings/rijsbergen79_infor_retriev.pdf

Apéndice

Apéndice 1. Pasos a seguir para el cálculo del Coeficiente de Dice.

1. Identificación de los bi-gramas únicos solapados de los términos a comparar. Por ejemplo, supongamos que extraemos de la noticia el término “arma” y que en EUROVOC en español aparece “armamento”. El script, al comparar ambos elementos determina que existen tres bi-gramas únicos compartidos.

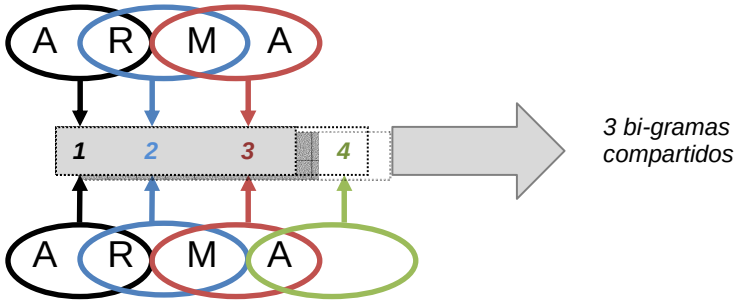

Figura 5. Identificación de bi-gramas solapados entre dos términos comparados.

2. Aplicación de la fórmula del Coeficiente de similitud de Dice (C_d) a cada palabra clave extraída y expresiones que integran los tesauros, en el mismo idioma. El resultado está comprendido entre 0 (ausencia de semejanza) y 1 (semejanza total): **C_d = (2 x C) / (A + B)**, donde A = Nº de bi-gramas únicos del 1º término; B = Nº de bi-gramas únicos del 2º término; C = Nº de bi-gramas únicos que comparten ambos términos. En el ejemplo el resultado es 0,75.

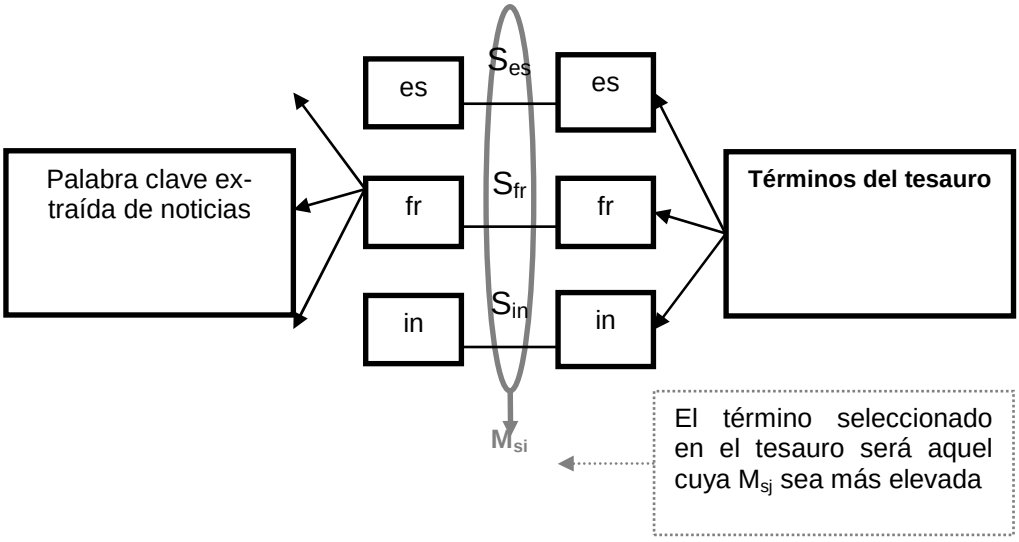

Figura 6. Cálculo del coeficiente de Dice de la palabra “arma”.

3. Cálculo de la media de los C_d para cada palabra clave extraída, en los tres idiomas. El término del tesauro (UNESCO u EUROVOC) seleccionado por el script será aquel con la media (M_{si}) de similitud más elevada: **M_{si}= (S_{es}+S_{fr}+S_{in}) / 3**, donde S_{es}= C_d para términos en español; S_{fr} = C_d para términos en francés; S_{in} = C_d para términos en Inglés. En el ejemplo, la Media del Coeficiente de Similitud de Dice = 0,783.

	Palabra clave	*Término en EUROVOC*	*Coeficiente Dice*
Español	Arma	Armamento	S_{es}= 0,75
Francés	Arme	Armement	S_{fr} = 1
Inglés	Weapon	Weapon	S_{in} = 0,6

Tabla 11. Cálculo del Coeficiente de similitud de Dice de la palabra clave "arma"