La traducción para indizar los contenidos de los recursos bibliográficos digitales

Catalina Naumis Peña
Centro Universitario de Investigaciones Bibliotecológicas, UNAM, México

Introducción

El objetivo del presente trabajo es destacar un tema recurrente en la literatura bibliotecológica actual que es el fenómeno de la indización temática multilingüe de los recursos bibliográficos digitales para recuperar información. Para lograr este objetivo se revisan artículos sobre el tema en los años transcurridos de este nuevo siglo y las soluciones que se proponen a través de la observación de un estudio de caso. Se analizó el caso del idioma chino porque es uno de los más abordados dentro del tema.

En alrededor de cuatrocientos años los índices han evolucionado para representar temáticamente las obras al interior de las propias obras o en las relaciones entre obras sobre asuntos específicos. Desde el índice de materias agregado a su libro *De scriptis medicis libri duo* en 1662 por Juan Antonio van der Linden (Malclés, 1960, p. 28) se han seguido refinando las técnicas de indización y sobre todo se han seguido reflejando en los documentos indizados los nuevos conocimientos que surgen.

Un índice no tiene valor separado de las obras que clasifica o tampoco en los registros de recursos bibliográficos, catálogos o bibliografías, porque

marca los puntos de acceso a las obras. La importancia del conjunto de elementos de acceso al documento lo explica Napoleón I en una carta fechada el 19 de abril de 1807, cuando anunciaba la creación de la École des Chartes "Si en una gran capital como París hubiese una escuela especial de historia donde se siguiera primero un curso de bibliografía, un joven, en vez de extraviarse durante meses en lecturas insuficientes o poco dignas de confianza, podría ir hacia las mejores obras y conseguiría más fácil y más rápidamente, mejor instrucción" (Malclés, 1960, p. 44). Después de doscientos años sigue siendo una frase con significado actual.

Desde la creación de la *École des Chartes*, pero sobre todo en la última década del siglo XX y en los años transcurridos del siglo XXI la humanidad ha experimentado grandes y aceleradas transformaciones económicas, políticas y culturales, y el presente trabajo refleja parte de esos cambios. La traducción y la indización realizadas ahora en forma automática son procesos que han evolucionado y muestran nuevos problemas que hay que resolver. La nueva revolución de la información o del conocimiento que se vive tiene consecuencias en las formas de trabajo, en la cultura y en la universalización del acceso a la información.

Al margen de este fenómeno social trascendente en cuanto a la transmisión del conocimiento no se puede olvidar la contrapartida del mismo: la gran comunicación existente y la facilidad de acceso a la información están asociadas a una honda desigualdad económica, una gran marginación social y un intenso deterioro ambiental. Parecería que a mayor información menor preocupación por el contexto en el que se vive y más individualismo de los seres y de las naciones.

Como siempre, existe una contrapartida de experiencias positivas que se reflejan en la mejoría de la calidad de vida de una buena parte de la humanidad: los países asiáticos aprovechan la información y el conocimiento propio y ajeno para desarrollar modelos de producción y desarrollo que los han ayudado a tener un éxito rotundo. No han pasado muchos años desde que las industrias automotrices japonesas enviaron alumnos a las fábricas automotrices occidentales a aprender y compraron coches de los mejores modelos producidos en el resto del mundo para copiarlos, hasta hoy, cuando los coches que más circulan en las calles del mundo occidental provienen del Oriente.

La producción que ha logrado tal éxito proviene de una fuerza de trabajo preparada, entrenada e impulsada por los diferentes conocimientos de un núcleo, que es el motor del logro obtenido. Este fenómeno observado a través de los acontecimientos de la vida cotidiana hace interesante el seguimiento de los trabajos que se realizan en sus sistemas de información que al fin y al cabo son los que proveen las fuentes de conocimiento que retroalimentan el proceso social de información. No sería adecuado analizar los sistemas de información a un nivel general porque toda observación debe partir del análisis de cada uno de los elementos que componen un sistema para finalizar con la suma de los resultados y evaluar el sistema en su totalidad.

Como el objetivo del presente trabajo representa sólo una pequeña parte del mundo de la información, en este caso concreto la indización temática multilingüe de los recursos bibliográficos digitales y las soluciones que se observan, se revisa la literatura bibliotecológica sobre sistemas de información en idioma chino para dilucidar si la indización temática también se encuentra entre los aspectos que preocupan y ocupan al mundo oriental, cuáles son los aspectos que se destacan de este fenómeno y cómo los solucionan.

La representación temática

La indización puede ser realizada con el modelo de los tesauros y/o las ontologías dependiendo del ámbito funcional en el cual se inserte. La diferencia del tesauro con la ontología es la estructura que se usa en uno y otro. Los dos están basados en mapas de conocimientos, el tesauro se enfoca en los temas de los contenidos documentales, y la ontología cumple funciones de diccionario para etiquetar con precisión el lenguaje de la Web, establecer sus propiedades y sus relaciones con otros conceptos, y transformarlos en términos para que cuando un software de búsqueda encuentre una palabra pueda interpretar su significado (Moreira González, 2004, p. 215). Estas herramientas lingüísticas han despertado un interés creciente porque apoyan la organización sistemática de la información mediante estructuras categorizadas de conocimientos y recopilan términos representativos de

ámbitos especializados (Gilchrist y Kivi, 2000; Hill y Koch, 2001; Hodge, 2000; Taylor, 2004; Tudhope y Koch, 2004; Williamson y Beghtol, 2003), (Roe y Thomas, 2004), citados por Caminotti y Martínez (2006, p. 75).

La organización documental presenta un conjunto de desafíos para adecuarse a los nuevos entornos tecnológicos y le brinda una respuesta oportuna a las necesidades de información de la sociedad. La información se transmite por diferentes medios de comunicación y llega a través de radio, televisión, periódicos, libros, revistas, Internet, discos, videos, i-pod y películas. Sin embargo tal cantidad de información no asegura que se obtenga conocimiento porque el proceso de absorción exige una activa respuesta humana. La transformación en conocimiento implica que se contraste una actitud analítica, sintética, crítica y reflexiva acerca de la información que se recibe, y que se contraste con la experiencia anterior y las necesidades actuales. Por ello, la información que proporciona el bibliotecólogo o documentalista deberá ser motivadora y disparadora de los procesos de conocimiento y no sólo una simple mediación pasiva, sino comprensiva de los documentos que se transmiten.

Para la Bibliotecología es un reto profundizar en la gran variedad de medios de transmisión del conocimiento y en los recursos bibliográficos que están ampliando su uso en la sociedad y por tanto exigen su representación en los sistemas de información. Los organismos internacionales de normalización bibliográfica están proponiendo nuevos sistemas de registro aplicando metadatos, que son la llave de acceso a los contenidos documentales y ello exige mayor profundización en los lenguajes de intercambio con las computadoras.

Los recursos bibliográficos digitales constituyen ceros y unos que exigen una organización que asegure la recuperación del documento original en su formato y con sus propias características y los elementos para su identificación a través de las bases de datos. La sociedad exige hoy con mayor fuerza que antes la organización de los medios digitales y esto es algo que el profesional del documento deberá hacer.

En las bibliotecas y servicios de información se controlan los servidores, las bases de datos y la Red misma, el almacenamiento de los

contenidos digitales, las aplicaciones para procesar y catalogar el contenido, y aquellas otras que buscan y recuperan el contenido. Otra tarea es la conversión y traslado a nuevos soportes tecnológicos para asegurar la conservación de los contenidos digitales. Pero lo más importante es implementar normas transparentes y protocolos que garanticen la interoperabilidad y compatibilidad de todos los ficheros y bases de datos en las bibliotecas y los sistemas de información.

Una buena organización del recurso bibliográfico digital es instrumentada por los metadatos que agregan características genéricas y abstractas para recuperar los contenidos documentales. Es decir los metadatos se agregan al contenido del documento sin formar parte del documento pero permiten manipularlo y relacionarlo con los otros documentos del sistema. En los metadatos el trabajo bibliotecario e informático están imbricados para transmitir la información y, por supuesto, los metadatos de contenido reflejan esta superposición del trabajo de una especialidad y otra. Son los metadatos los que nos permiten relacionar los documentos que contienen los mismos temas en forma inmediata, aunque se encuentren en diferentes idiomas, y es aquí donde se produce una simbiosis disciplinar que debemos analizar desde la bibliotecología para definir la participación que nos corresponde, aunque el alcance de este trabajo es únicamente exploratorio del tema y sólo ofrece una opinión.

Los lenguajes documentales, como los tesauros, definen, estructuran, desambiguan y relacionan los términos en forma previa a su ingreso a los sistemas de información, pero esos tesauros se pueden usar adaptándolos en el medio digital, convirtiéndolos en ontologías. Es decir, se usan los términos y sus relaciones, pero se establecen nuevos procesos informáticos que ligan términos entre el sistema lingüístico y el sistema de información al momento de hacer la indización y las agrupaciones temáticas de documentos para recuperarlos.

Uno de los grandes problemas que hay que resolver para indizar y representar los contenidos documentales a nivel mundial es la conversión entre idiomas. La tecnología abate las distancias entre países y permite intercambiar la información generada, pero la comunicación científica debe vencer otras barreras, como la lingüística. La investigación y el desarrollo de colecciones digitales incluyen la creación de

contenidos, la conversión, la indización, la organización y la diseminación, y todos y cada uno de estos procesos están ligados y hacen posible la recuperación de los contenidos digitales y además pueden ser sometidos a procesos de traducción de un idioma a otro. Hay casos de traducción que son más o menos sencillos porque existen muchas referencias culturales, el nuevo reto son los idiomas que se caracterizan por importantes diferencias en sus estructuras morfológicas, incluso a veces opuestas. La creación de sistemas de información globales supone el uso de herramientas capaces de procesar e intercambiar conocimientos generados en cualquier idioma. Hace falta entonces un trabajo conjunto entre traductores, bibliotecólogos e informáticos para producir índices multilingües que acerquen el conocimiento a un ámbito global.

Los términos de indización documental

Cuando se estudia la representación temática en los sistemas de información se están resolviendo problemas de comunicación, un término mal empleado en la indización o en la expresión usada por un autor lleva a la ausencia de información, o a lo que se conoce como silencio informativo, porque si el usuario busca con el término que tiene significado para él y el sistema no tiene éste registrado, no recuperará los documentos que están representados con un término sinónimo no conocido por ese usuario. La ambigüedad de los términos debe ser informada a los usuarios, pero los términos de entrada a los sistemas deben reflejar los usos lingüísticos de una cultura y eso es lo que resulta difícil de traducir.

Por ejemplo un documento usa la expresión AGOTAMIENTO DEL AGUA que en México es designada como ESCASEZ DE AGUA, y así cuando un usuario mexicano busca por ESCASEZ DE AGUA en un sistema de información en español, no siempre tendrá en mente que ese sistema puede haber elegido una expresión distinta como AGOTAMIENTO DEL AGUA. Para asegurar que el usuario cuente con recordatorios de equivalencias, el sistema de información tiene que estar preparado para ello, y la solución más común para recuperar con equivalencias es

disponer de una herramienta lingüística como un tesauro del cual se extraigan los términos de indización y que cuentan con los términos equivalentes que pueden ser usados en la comunicación. De este modo se podría recuperar el mismo documento usando cualquiera de los dos términos, sin un esfuerzo adicional del usuario.

Existen otros ejemplos como el uso de AGUAS BLANCAS en lugar de AGUA DE LLUVIA o de AGUAS NEGRAS en lugar de AGUAS RESIDUALES. En estos casos la primera palabra del término es la misma y puede ayudar a recuperar los términos que no son adecuados. Sin embargo un término como REPRESAS para referirse a lo que en México se conoce como PRESAS (muro construido a través de un río, con objeto de regular su caudal o embalsar agua para aprovecharla en el riego o la producción de fuerza hidráulica) puede resultar más confuso para los usuarios. En este último caso el término usado en México mantiene un sentido polisémico en el español de otros países y resulta importante mantener una equivalencia entre las dos expresiones para que la interpretación del término no vaya a la deriva al intercambiar información con un público no especializado.

Continuando con otros ejemplos existe un término del español que crea conflicto y es JACINTOS, mejor conocidos en México como LIRIOS ACUÁTICOS. En todos estos casos se observa la necesidad de hallar equivalencias entre los términos para guiar al usuario en su búsqueda (Vargas Suárez, 2007, pp. 76).

Cuando se realizan las traducciones a otros idiomas estos problemas no son menores, porque deben analizarse para cada término, las correspondencias semánticas en el idioma que se traduce y al que se traduce. Resolver las traducciones de los términos especializados es uno de los procesos a los que destinan grandes recursos las grandes bases de datos y que se reflejan en la literatura sobre organización de la información. En el lenguaje general las conversiones de un idioma a otro quizás no resulten excesivamente problemáticas porque existen diccionarios de sinónimos que aclaran los diferentes usos de las palabras; sin embargo, cuando se trabaja con un lenguaje especializado la situación cambia radicalmente. No siempre existen terminologías en las diferentes especialidades o diccionarios de sinónimos que aclaren los significados.

Evidentemente las traducciones de términos especializados deben ser cuidadas y buscar la conversión a términos que tengan el mismo significado cultural en el idioma al que se deben traducir. La revisión de la literatura muestra una preocupación intensa en este aspecto, pero se advierte una tendencia muy fuerte por resolver la conversión mediante sistemas automáticos los cuales hacen hincapié en la seguridad y economía que ofrecen. Si bien la autora considera esta solución cuestionable, el presente trabajo sólo busca rescatar las diferentes propuestas que requerirán un análisis posterior.

Desde hace treinta años, con la popularización de las grandes bases de datos como el Chemical Abstracts, las comunidades científicas consultan en inglés y por lo tanto las unidades de información de todo el mundo les pagan a las empresas que distribuyen estas bases. Las grandes bases de datos ofrecen sistemas automáticos de traducción de los documentos en su sistema y se han incrementado los trabajos que tratan de resolver los problemas de traducción, tanto en cuanto a los términos de indización, como a los del propio documento. Desde la aparición de Internet y las páginas WEB, la situación ha ido cambiando y las páginas de Internet en otros idiomas se han ido incrementando y los países están distribuyendo mayor cantidad de información en sus propios idiomas. Por eso la conversión terminológica entre idiomas es un gran tema en la organización de la información, porque en un texto completo la concatenación de las frases ayuda al entendimiento. La representación documental a través de palabras claves que indiquen el contenido de un texto en diferentes idiomas supone no sólo el conocimiento de los idiomas en los que se debe trabajar, sino el estudio de las solicitudes y expectativas de los usuarios, y por supuesto el análisis textual para entenderlo.

Para traducir un lenguaje en otro, los informáticos aprendieron que tenían que entender la sintaxis de ambos lenguajes, al menos en el nivel morfológico (la sintaxis de las palabras) y las frases enteras. Pero para entender realmente o a fondo la sintaxis, se debe entender la semántica del vocabulario y la pragmática del lenguaje. Así, lo que empezó como un esfuerzo para traducir textos se convirtió en la investigación para entender cómo representar y procesar el lenguaje natural usando computadoras.

Sin embargo cuando se habla de la traducción entre idiomas como el chino y el inglés que difieren ampliamente en las estructuras lingüísticas y en la organización de los caracteres, los retos parecen mucho mayores que los observados entre usos diferentes en el mismo idioma o idiomas con bases similares como el inglés y el español.

La conversión del lenguaje de indización al chino supone en general la traducción de nombres de personas, instituciones, monumentos y nombres propios. La indización no se restringe a los problemas de significado de los contenidos documentales, sino que comprende los diferentes elementos de representación documental, además de los temáticos. De hecho éste es uno de los problemas que se menciona con mayor frecuencia en la literatura sobre representación multilingüe, porque los sistemas de traducción trabajan sobre diccionarios que hacen la conversión y los nombres propios no están incluidos en los diccionarios que realizan las traducciones en los sistemas de información.

Investigaciones sobre representación temática reportadas en la literatura bibliotecológica en chino

La literatura bibliotecológica reporta investigaciones en materia de representación temática en muchos países, pero se destaca una literatura abundante proveniente del bloque asiático para abatir las barreras lingüísticas y recuperar información en otros idiomas. En 1999 se hicieron estudios que preveían un 150% de crecimiento de contenidos en otros idiomas que no fueron el inglés en la Red, para los siguientes próximos cinco años. Un reporte del año 2000 mostraba que los usuarios de Internet en chino crecieron de 8.9 millones a 16.9 millones entre enero y junio de ese año. En el 2002 el chino se convirtió en la segunda lengua en Internet y ha seguido incrementándose. En el 2007 tuvo un crecimiento del 469.6%[1] y desde el 2000 al 2008 ese crecimiento fue del

1 "Número de usuarios de Internet por lenguas", Miniwatts Marketing Group, 2000-2007, disponible en http://www. Internetworldstats.com/stats7.htm (Consultada el 11 de noviembre de 2007).

755.1% y los usuarios en chino constituyen el 20.4% del porcentaje de usuarios de Internet.

La muy fuerte entrada de los orientales como nuevas potencias económicas viene acompañada de una preocupación por conocer la información que generan los hablantes en otros idiomas. El chino es la lengua oficial en China, Taiwán y Singapur, pero también se habla en Indonesia y Malasia. En realidad, se trata más bien de una familia de lenguas que de una sola lengua, aunque la tradición cultural china prefiere llamar dialectos a sus múltiples variedades lingüísticas. El chino es el idioma más hablado del mundo y es uno de los seis idiomas oficiales de la Organización de las Naciones Unidas.

El sistema de la lengua china es mixto, está compuesto por ideogramas y sonidos fonéticos. Pero existe un elemento de suprema importancia en la escritura china, que son los radicales; también denominados 'claves', 'clasificadores' o 'determinativos'; los radicales son los determinantes que indican el significado de los caracteres. Son éstos los ladrillos de la escritura china, pues cada carácter chino *(hànzí)* es o un radical o un compuesto que tiene de dos a cuatro radicales. Por medio de estos radicales se organizan los diccionarios. Las nuevas tecnologías informáticas han permitido escribir y leer el chino con mucha facilidad a pesar de ofrecer ésta muchas dificultades porque no existen límites claros entre las palabras y los ideogramas cambian de significado dependiendo de la posición en que se encuentren en una frase.

En el medio informático relacionado con bases de datos se hace mucho hincapié en la importancia que tiene una buena indización para que la búsqueda sea exitosa. Se dice además que existen muchos trabajos en inglés y muy pocos en chino (Yang, Luk, Yung & Yen, 2000, p. 341). Por tanto las propuestas de indización automática se enfocan en buena medida en la indización en chino, empleando como medio la conversión del chino al inglés y del inglés al chino. El estudio de las conversiones está ligado además al uso de dominios semánticos y del conocimiento, que es otro de los aspectos recurrentes en la investigación sobre la recuperación de información. Cada uno de los reportes de investigación que se comentan a continuación está ligado a un dominio semántico en particular.

En la literatura bibliotecológica siempre se ha insistido en que el proceso de indización debe estar sustentado en la práctica desde la cual se instrumentan los descriptores y el estudio teórico para elaborar los mapas conceptuales que representan la estructura del conocimiento subyacente en los sistemas de información. Charles A. Cutter definió el *principio del uso* como uno de los fundamentos del vocabulario controlado (Foskett, 1982), y durante muchos años este principio fue interpretado como el término más correcto en el lenguaje escrito; sin embargo en la comunicación actual se prefiere el término más habitual entre los usuarios finales. Las máquinas trabajan sobre los términos que establecen los propios autores, porque su fuente de observación y su rescate de los términos de indización son los documentos que están en los sistemas de información. Los motores de búsqueda rescatan además los términos usados en la consulta a los sistemas, pero los sistemas de recuperación en idioma chino hacen énfasis en las solicitudes de los usuarios hacia el conjunto documental; es decir la representación del contenido rescatada a través de las preguntas de quienes consultan (Bian & Chen, 2000, p. 282).

Uno de los trabajos publicados y desarrollados en Taiwán explica los métodos desarrollados para implementar sistemas de recuperación que hacen la traducción de las preguntas de las búsquedas a los términos que usan los sistemas. Se usan las preguntas en un medio chino para acceder a información en chino e inglés. Tanto las preguntas de los usuarios como las respuestas del sistema están en chino, sin importar la lengua de los documentos recuperados. Las traducciones de documentos del chino al inglés se realizan a partir de un diccionario bilingüe y un corpus monolingüe en inglés, de donde se seleccionan los términos más adecuados. Uno de los problemas más frecuentes con respecto a la traducción es la celeridad en la creación de nuevos términos que no se alcanzan a incorporar al diccionario. El otro es que en chino las frases están compuestas de una serie de caracteres sin límites precisos.

Los autores destacan que los porcentajes de las preguntas a los sistemas de información que contienen nombres propios son muy altos. Mencionan el trabajo de Thompson & Dossier quienes analizaron las preguntas hechas a tres periódicos americanos en unos cuantos días

de 1995, y los porcentajes que incluían nombres en las preguntas fueron de 67.8, 83.4 y 38.8. Los autores utilizaron algoritmos para transliterar nombres propios, y con el apoyo de estos sistemas se tradujeron cerca de 100.000 páginas WEB en los últimos meses de 1997 (Bian & Chen, 2000, pp. 281-296).

Otro de los grandes proyectos que incluye la conversión de la indización en chino es realizado en Estados Unidos de Norteamérica y financiado por la National Science Foundation que apoya soluciones para búsquedas en la Red. Esta financiación está inscrita en un proyecto más amplio que es la National Information Infrastructure. La investigación y el desarrollo de bibliotecas digitales incluye la creación de contenidos, la conversión, la indización, la organización y la diseminación. La clave del desarrollo tecnológico es "cómo buscar" y cómo desplegar lo seleccionado desde y a través de grandes colecciones. Los proyectos de investigación sobre bibliotecas digitales financiados por la National Science Foundation de los Estados Unidos tienen como tema común brindar búsquedas en la Red, que es la bandera de investigación de la National Information Infrastructure. La promoción de este trabajo se realiza bajo la consigna de que la indización es una tarea importante para recuperar información y que los mejores resultados dependerán de indizaciones bien hechas, y su propósito último es que se debe desarrollar una biblioteca digital universal para satisfacer el sueño acariciado por muchos investigadores.

En este sentido el problema más crítico es recuperar a través de un idioma documentos que están en otros idiomas o hacer esto entre bibliotecas digitales multilingües. El trabajo se enfoca a la indización en chino pero la intención profunda es recuperar información de documentos en otros idiomas. En las bibliotecas digitales chinas la unidad más pequeña de indización son las palabras, mientras que en una oración en chino la unidad más pequeña son los caracteres que se suman a la característica mencionada de que los textos chinos no tienen marcas de comienzo y terminación de una palabra, como los textos en inglés. Sin embargo los textos chinos no tienen delimitadores que marquen donde comienza y termina una palabra, como los textos en inglés. En el inglés, como en otros lenguajes a menudo que utilizan ortografía romana o griega, los espacios indican los límites

de las palabras. La escritura china consiste de cadenas de caracteres (o ideogramas) separadas por puntuación y las traducciones se realizan a partir de diccionarios bilingües. Este proyecto trata de resolver mediante métodos estadísticos el reconocimiento de las expresiones que no están en los diccionarios, como nombres, lugares, eventos y temas especializados en un sistema de información que, al igual que el proyecto anterior, indiza periódicos (Yang, Luk, Yung & Yen, 2000, pp. 340-351).

El trabajo que se comenta a continuación está relacionado con búsquedas de información efectuadas en bases de datos de Taiwán. Se parte de la base que las solicitudes de búsquedas en la Web contienen un promedio de 2.3 palabras, que en realidad son pocas para expresar un interés que coincida con las expectativas de los usuarios. Se trata de un desarrollo que sugiere términos para la búsqueda interactiva en la Web que se propone ayudar a los usuarios a formular mejores preguntas para reducir el ámbito de búsqueda, y además efectúa sugerencias de términos extraídos de descriptores co-ocurrentes, de acuerdo con su frecuencia de uso. Estos desarrollos se enfrentan a múltiples problemas, tales como la interferencia de documentos irrelevantes, así como a la extracción de términos relacionados conceptualmente, pero que no son co-ocurrentes en los documentos.

El nuevo desarrollo propuesto permitirá extraer términos relevantes de varias búsquedas realizadas por usuarios diferentes para identificar sus demandas y la aproximación a documentos en los que coinciden algunos términos que permiten sugerir términos de recuperación más relevantes. En la utilización de este sistema, los términos relevantes sugeridos por los usuarios en las preguntas, serán co-ocurrentes, al igual que los términos sugeridos por el sistema en las casillas de búsqueda. Adicionalmente, los términos sugeridos en cada búsqueda interactiva pueden ser organizados, de acuerdo con su relevancia, en una sesión completa de búsqueda, con mayor rapidez que una búsqueda con un desarrollo convencional.

El paquete de programas se probó utilizando un servidor *proxy* con una casilla de búsqueda que contiene dos millones de transacciones de preguntas enviadas a buscadores en Taiwán. El resultado experimental mostró que el desarrollo puede organizar la información en

términos bastante relevantes, y explotar la información en su contexto durante las sesiones de búsqueda, lo que le permitirá hacer sugerencias efectivas (Huang, Chien & Oyang, 2003, pp. 638-649).

El proyecto que se menciona a continuación trata de imitar el proceso que realiza el indizador humano al comparar los términos de un documento y los contenidos en el vocabulario para traducirlos a la terminología definida en un determinado sistema de información. Aquí el proceso se realiza automáticamente y los resultados deberán ser observados en la práctica, ya que se propone un modelo de recuperación de información conceptual basado en la navegación en dos tipos de diccionarios: un diccionario global y uno local. El diccionario global con términos autorizados se utiliza en la captura de los conceptos comúnmente conocidos relacionados con la pregunta y el documento, y reemplazar las palabras claves con términos del diccionario. Los documentos son clasificados según la cercanía conceptual con la pregunta, y son organizados en un formato para biblioteca digital personal, pDL (personal Digital Library), destinado a los usuarios. El usuario puede navegar en los documentos pDL sugeridos por la biblioteca digital y examinarlos exhaustivamente. Esta sugerencia se hace a través de la información en un diccionario local que se organiza de tal manera que refleje los intereses de los usuarios y la asociación de las palabras claves con los documentos. En los experimentos para mejorar la recuperación, se usaron los dos tipos de diccionarios desarrollados con colecciones estandarizadas (Nakayima, Sato, Qu, & Ito, 2003, pp. 16-28).

Se reporta además otro estudio que presenta un tesauro elaborado por medios automáticos para recuperar información legal generada en el Departamento de Justicia de Hong Kong. En Hong Kong el inglés y el chino son idiomas oficiales, pero como también se destaca en el proyecto anterior, los dos idiomas tienen diferencias significativas en su estructura y gramática. Por medio del tesauro que convierte los términos del chino al inglés y al revés se brinda la oportunidad de conocer las decisiones judiciales desde otro idioma que no es el de origen del documento. Primero se realizó la extracción de los términos significativos de los mismos documentos en los dos idiomas para contar con un corpus paralelo comparable. El tesauro fue generado automáticamente a través de la conversión entre los dos idiomas utilizando

el análisis de co-ocurrencias y la red de Hopfield, que es una técnica informática de contenido direccionable de la memoria (Yang & Luk, 2003, pp. 671-682).

La seguridad nacional requiere de una gran cantidad de datos e información que se generan y recopilan diariamente, y muchos de estos datos están escritos en lenguajes diferentes, almacenados en diferentes lugares y probablemente no entre sí. La interoperabilidad semántica en la conversión de lenguaje es un gran cambio que se genera al revisar la información y los datos dispares, los cuales hay que analizar, mostrar, buscar y resumir. Los ataques terroristas del 11 de septiembre de 2001 han incrementado la atención de la seguridad nacional y el análisis criminal, países asiáticos como Japón, Taiwán y Singapur han sido advertidos que podrían ser los siguientes blancos de ataques terroristas. Por esta razón la investigación en recuperación de información sobre el tema de seguridad se ha enfocado en la interoperabilidad semántica y, como consecuencia, en la construcción de tesauros.

El desarrollo tradicional de la recuperación de información requiere normalmente un documento que muestre algunas palabras clave incluidas en la pregunta. La generación de asociaciones para términos relacionados entre dos espacios de usuarios y documentos es importante y se vislumbra la solución del problema con la creación de un tesauro. La ambigüedad en la traducción incrementa significativamente el problema de la recuperación y se hace extensiva a la interoperabilidad semántica en la conversión de lenguaje.

El trabajo se enfoca al problema de la interoperabilidad semántica en la conversión del inglés al chino; sin embargo, el desarrollo de técnicas no se limita a estos lenguajes y puede ser aplicado a otros idiomas, aunque obviamente el inglés y el chino son los idiomas que interesan en esta región asiática. Mucha información sobre seguridad nacional y criminal se encuentra en estos idiomas. La generación de un eficiente tesauro automático entre estos idiomas es importante para recuperar información entre el inglés y el chino. Para facilitar la conversión de un lenguaje con la finalidad de recuperar información se desarrolló un *corpus* que usa términos estadísticamente co-ocurrentes, y a la vez un cuerpo que compara y construye un modelo de conversión de lenguaje, que al mismo tiempo realiza la traducción de los términos.

El estudio se desarrolló a partir de los documentos solicitados en la web a través del Departamento de Policía de Hong Kong en los idiomas inglés/chino, pero también se introdujo un algoritmo que genera una base robusta de conocimiento basada en correlaciones de análisis estadísticos de las relaciones de significado, que se integra al presionar la opción bilingüe del corpus. La investigación de los resultados permite evaluar cómo un tesauro basado en una red de conocimiento semántico producto de un análisis estadístico, puede ayudar en la conversión del lenguaje utilizado en la información y su recuperación (Li & Yang, 2005, pp. 272-282).

Un último proyecto de traducción que busca indizar en chino es el siguiente, en el cual se propone un esquema que facilite la recuperación de imágenes en determinado dominio de conocimiento basado en una ontología y un tesauro. En particular se basa en la enseñanza del sistema basada en casos que utilizan frases en lenguaje natural transferidas a un lenguaje de programación, las cuales son propuestas como preguntas dentro de un formato RDF. Este mismo lenguaje de programación también se extiende al desarrollo de anotaciones semánticas que describen los metadatos de imágenes y los convierten automáticamente al mismo esquema RDF. La recuperación de imágenes puede conducirse por la construcción de estructuras semánticas usadas en las preguntas de la búsqueda, con los metadatos antes descritos. La búsqueda se efectuó en un dominio de Historia, recuperando imágenes históricas y culturales tomadas del disco compacto del Doctor Ching-chih Chen "First Emperor of China", como parte de una colaboración internacional en bibliotecas digitales. Se construyó e implementó la ontología del dominio, un tesauro en chino-mandarín, así como la construcción y recuperación de algoritmos con la finalidad de probar el sistema propuesto (Soo, Lee, Li, Chen & Chen, 2003).

Consideraciones finales

Los elementos de indización que hay que traducir en los sistemas de información chinos son además de los términos que representan los contenidos documentales: nombres de personas, instituciones, monumentos y nombres propios en general

En cuanto a la asignación de la indización temática, los estudios presentados en este trabajo fueron desarrollados sobre documentos digitalizados, generados en su mayor parte sin un análisis documental por asignación, para proponer el lenguaje disciplinar que serviría de base a la indización y la traducción. La bibliotecología ha realizado a través de los años la indización como un proceso previo a la incorporación de documentos a los sistemas de información. Las herramientas lingüísticas que se generaban para realizar la indización desde la bibliotecología contenían un corpus lingüístico que representaba un modelo a escala de la variedad de la lengua en un determinado ámbito del conocimiento. Las traducciones eran hechas sobre el lenguaje establecido en el modelo.

En ese sentido existe una definición de *corpus* que precisa el concepto de representatividad, que es el "equilibrio conceptual" o "conceptual balance" el cual supone recoger todos los subcampos en los que se divide cada parte del conocimiento, así como todos los ámbitos especializados que guardan relación con ésta (si es que se trata de un ámbito multidisciplinar) De acuerdo con este concepto, parece más fácil que el *corpus* sea representativo si incluye todas las variedades textuales típicas del área objeto de estudio y aborda todos los subcampos en los que ésta se divide y aquellos con los que guarda relación. El concepto de *equilibrio conceptual* fue introducido por Bowker (1999, p. 45).

En general en los artículos revisados no se plantea la elaboración de una herramienta lingüística basada en los sistemas de conocimiento para organizar la información de las colecciones de recursos digitales o documentos, sino una recopilación de términos usados en los documentos y en las preguntas de los usuarios a los sistemas. Es decir el desarrollo informático busca definir la situación comunicativa, que en cada caso tiene una naturaleza diferente (Chang & Zen, 2002). La tendencia es atender de manera puntual cada situación que se presenta, sin buscar la generalización de las soluciones deseables en un futuro para los sistemas de información. Los problemas abordados, excepto en la indización de imágenes, están relacionados con la indización automática, ya sea de temas o de entidades personales o corporativas.

Con excepciones, en los trabajos analizados prevalece la preocupación por la traducción del idioma chino al inglés y del inglés al chino

más que una organización temática para indizar y representar un área de conocimiento. El acceso a la información de un lenguaje a otro está enfocado a resolver la traducción de las preguntas y los problemas de traducción de los documentos.

Según los propios autores los resultados de la traducción no permiten captar totalmente la estructura de la información, en los casos que se trabaja con la pregunta en chino y los contenidos de los documentos en chino y en inglés, se obtiene un porcentaje de aceptación de la traducción de un 67.74%.

En general en los proyectos propuestos solo se recuperan los documentos en los que existe coincidencia terminológica, pero se pierden aquellos en los que existe equivalencia.

La representación temática es abordada con frecuencia a través de las preguntas formuladas por los usuarios de los sistemas y las palabras usadas por los autores de los documentos para nombrar los conceptos que tratan de transmitir, y no producto de la reflexión sobre cada uno de los contenidos documentales, los conocimientos de un área de estudio y los intereses de los usuarios de un sistema, que son los tres elementos básicos para diseñar una indización eficiente.

Las preocupaciones por la conversión automática de los nombres de las entidades ocupan un lugar importante dada la diferente estructura del chino y del inglés.

En otro caso se estudia la mejoría de los términos de búsqueda que se le sugieren al usuario de un sistema de información basándose en la comparación de las preguntas de diferentes usuarios. Como se puede observar es un estudio de indización automática de las preguntas que se le hacen al sistema y en este caso es probable que exista algún margen de equivalencia terminológica.

La excepción a los estudios mencionados hasta aquí es la construcción de tres tesauros documentales que se obtuvieron como resultado de operaciones automáticas: uno sobre un corpus legal paralelo generado por el Departamento de Justicia de Hong Kong para recuperar decisiones legales en chino y en inglés; y el otro tesauro se construyó para apoyar un sistema de información policial sobre seguridad. Llama la atención que estos dos proyectos hayan estado destinados a temas legales que requieren de gran precisión en el lenguaje utilizado. A diferencia

de otros proyectos, las preguntas de los usuarios y las respuestas de los contenidos documentales fueron abordadas a partir de un cuerpo de conocimientos: los tres elementos mencionados en párrafos anteriores. El tercer tesauro fue desarrollado en una base de datos sobre imágenes y cuenta con el apoyo que supone el uso de expertos para definir temas sobre un conjunto de documentos. Se definieron los contenidos de las imágenes a través de un tesauro y una ontología en chino-mandarín. En la indización de imágenes es éste el único caso de la literatura revisada en que se reconoce la importancia de que la calificación haya estado a cargo de un ser humano.

En general la indización de sistemas de información traducidos a partir de otros idiomas que se propone en los trabajos analizados es de tipo automatizada, pero con frecuencia se usan colecciones de documentos sistematizados por documentalistas y archivistas, y es aquí donde se podrán hacer estudios que permitan definir las diferencias entre la conversión idiomática sobre colecciones organizadas con indización por asignación para recuperar sus contenidos y las que no lo están. Por último es necesario destacar que en los trabajos de traducción de índices chino-inglés se usa el término tesauro y no ontología, a pesar que se refieren a tesauros cuyos términos reciben un tratamiento informático para sus relaciones.

Obras consultadas

Bian, G. W. & Chen, H. H. 2000, "Cross-language information access to multilingual collections on the Internet", en *Journal of the American Society for Information Science and Technology,* vol. 51, núm. 3, pp. 281-296.

Bowker, G. C. & Star, S. L., 1999, *Sorting things out: classification and its consequences,* Cambridge, Mass, MIT.

Caminotti, M. L. & Martínez, A. M., 2006, "Fútbol, tesauros y taxonomías WEB: desafíos del control del vocabulario", *Información, Cultura y Sociedad,* núm. 14, pp. 73-81.

Chan, L. M. & Zeng, M. L., 2002, "Ensuring interoperability among Subject Vocabularies and Knowledge Organization Schemes: a methodological analysis", IFLA COUNCIL AND GENERAL CONFERENCE (68: agosto 18-24, 2002: Glasgow).

Foskett, A. C., 1982, *The subject approach to information,* 4th ed, London, Clive Bingley.

Huang, C.K., Chien, L. F. & Oyang, Y. J., 2003, "Relevant term suggestion in interactive Web search based on contextual information in query session logs", en *Journal of the American Society for Information Science and Technology,* vol. 54, núm. 7, pp. 638-649.

Li K. W. & Yang, C. C., 2005, "Automatic Crosslingual Thesaurus Generated from the Hong Kong SAR Police Department Web Corpus for Crime Analysis", en *Journal of the American Society for Information Science and Technology,* vol. 56, núm. 3, pp. 272-282.

Malclés, L. N., 1980, *La bibliografía,* Buenos Aires : EUDEBA.

Nakashima M, Sato, K., Qu, Y. & Ito, T., 2003, "Browsing-based conceptual information retrieval incorporating dictionary term relations, keyword association, and a user´s interest", en *Journal of the American Society for Information Science and Technology,* vol. 54, núm. 1, pp. 16-28.

Soo V. W., Lee, C. Y., Chung-Cheng Li, Chen, S. L., & Chen, C. C., 2003, "Automated semantic annotation and retrieval based on sharable Ontology and case-based learning techniques", en *Proceedings of the 2003 Joint Conference on Digital Libraries,* IEEE Computer Society.

Vargas Suárez, V. E. 2007, "Análisis del lenguaje relacionado con el agua: necesidad de un tesauro para México", Tesis de Maestría, UNAM, Posgrado en Bibliotecología y Estudios de la Información.

Yang C. C. & Luk, J., 2003, "Automatic generation of English/ Chinese Thesaurus based on a parallel corpus in Laws", en *Journal of the American Society for Information Science and Technology,* vol. 54, núm. 7, pp. 671-682.

Yang C. C., Luk, J., Yung, S. K. & Yen, J., 2000, "Combination and boundary detection approaches on Chinese indexing", en *Journal of the American Society for Information Science,* vol. 51, núm. 4, pp. 340-351.