OCENET Consulta: un portal de consulta en español basado en tecnologías lingüísticas

INTRODUCCIÓN

El presente estudio describe cómo el Grupo Editorial Océano ha desarrollado un nuevo portal de consulta general muy útil para estudiantes y educadores de habla hispana, denominado *OCENET Consulta* (http://ocenet.oceano.com), que ofrece un gran volumen de información, estructurada y publicada según estrictos criterios editoriales, a la cual se accede mediante un potente motor de búsqueda que utiliza diversas soluciones de ingeniería lingüística desarrolladas por CliC/Thera – *Centre de Llenguatge i Computació* (Universidad de Barcelona) .

PERFIL DE LAS EMPRESAS

Editorial Océano (http://www.oceano.com) es uno de los grupos editoriales más importantes del mundo de habla hispana, con más de 100 años de historia. La empresa, con sede central en Barcelona, está establecida en 21 países, entre los que se cuentan España, Portugal, EE.UU y todos los países de América Latina, y está incrementando su presencia en Europa del Este.

Aunque la andadura de Grupo Océano se inicia en 1959 con la fundación de Ediciones Danae, la posterior adquisición del prestigioso fondo de Instituto Gallach de Librería y Ediciones, que había sido fundado en 1899, permite asociar el nombre de Océano a una tradición editorial de más de un siglo. Coronando una trayectoria ascendente en todos los ámbitos de la edición y en todos los canales de distribución y venta, se crea en 1994 el sello Oceano Multimedia, especializado en los nuevos soportes electrónicos, y Oceano Digital, que aparece en 2002 y se ocupa específicamente de la creación de contenidos para Internet.

CliC/Thera (http://clic.fil.ub.es), es una empresa (*spin-off*) de la Universidad de Barcelona que orienta su actividad a la incorporación de los avances en ingeniería lingüística en los sectores productivos, especialmente en el sector de las comunicaciones, de la edición y la enseñanza. Las principales áreas de actividad de CliC/Thera son las siguientes:

- Transferencia de Tecnología de la Universidad al mundo empresarial, desarrollando aplicaciones que integran módulos de procesamiento del lenguaje así como recursos de ingeniería lingüística para su incorporación a sectores productivos relacionados con las Industrias de la Lengua y la Sociedad de la Información
- Impulso de la investigación en el ámbito de la ingeniería lingüística y la gestión del conocimiento
- Definición de nuevos perfiles profesionales en el campo de los servicios, los recursos y la ingeniería lingüística

OCENET CONSULTA: UN PORTAL PARA ESTUDIANTES, EDUCADORES Y PARA EL PÚBLICO GENERAL DE HABLA HISPANA

OCENET Consulta es una base de datos documental en Internet que constituye un centro de información integrado por contenidos de referencia, revistas y fuentes primarias que cubren diversas áreas de conocimiento a las que se accede desde un único punto.

OCENET Consulta está orientado a todo el público y en especial a los profesores y estudiantes de habla hispana que necesiten recopilar información fiable de una manera ágil y eficaz. El servicio permite realizar una búsqueda de información en distintos tipos de recursos (obras de referencia, revistas, textos primarios, biografías, cronologías, fotografías, diccionarios, etc.) y seleccionar aquellas categorías que más interesan a la persona que realiza la consulta (entradas de enciclopedias, artículos de revista, biografías, entradas de diccionarios, fotografías, información histórica, etc. Pero definitivamente la gran ventaja de del sistema de recuperación de *OCENET Consulta* es que posibilita un resultado avanzado al usuario no experto.

Los contenidos de *OCENET Consulta* corresponden a los niveles propios de la escuela primaria, secundaria y de primer ciclo universitario, por lo que los perfiles de usuarios a los que va dirigido este portal son estudiantes de bachillerato y universitarios de primer ciclo, así como a personas interesadas en obtener información general y de referencia. Los principales clientes de *OCENET Consulta* son redes de bibliotecas y universidades, que operarán como los puntos de acceso al portal para profesores y estudiantes.

Está previsto que este portal, cuya primera versión comercial salió al mercado el día 30 de octubre de 2002 como servicio de pago mediante suscripción anual, alcance a finales de 2003 la cifra de 3.000 usuarios, que accederán al servicio a través de los 150 clientes (fundamentalmente bibliotecas) que se espera haber obtenido para dicha fecha.

Editorial Océano señala que las principales características diferenciales del portal son que "la información suministrada está sujeta a un proceso editorial riguroso y que dispone de un motor de búsqueda muy eficaz"; este motor utiliza un sistema de indexación y descriptores y opera sobre la estructura XML de cada artículo para mostrar los resultados más relevantes, permitiendo estructurar la información según diversos criterios gracias al uso de descriptores. Por ello, es posible encontrar la información de las siguientes maneras:

- Por relevancia según los descriptores y la estructura de cada artículo
- Por tema (geografía, historia, etc.)
- Por tipo de contenidos (biografías, cronología, diccionarios, imágenes, etc.)

La base documental de *OCENET Consulta* consta de información de referencia, artículos de revistas y fuentes primarias. La información de referencia procede, en su mayoría, de la base documental de Océano, aunque también contiene aportaciones de otras editoriales que colaboran en el proyecto, como obras de GALE y de Bompiani, o revistas como *Clarín*, *Chasqui*, *National Geographic*, *Quimera*, *Realidad*, hasta alcanzar un total de sesenta de

publicaciones, seleccionadas entre las más relevantes en cada especialidad, y que se pretende que sean cien en la versión 2.0 programada para octubre de 2003.

La edición inicial incluye más de 300.000 páginas digitalizadas y más de 150.000 artículos. Finalmente, las fuentes primarias incluidas en *OCENET Consulta* (más de 2.000 documentos) son reproducciones digitalizadas e indexadas de documentos originales como, por ejemplo, el *Diario del primer viaje de Cristóbal Colón*, así como de obras clásicas de la literatura como pueda ser *El Quijote*.

Editorial Océano menciona que la integración de tecnología lingüística a los desarrollos, contenidos y soluciones tecnológicas de *OCENET Consulta* permite que este servicio se diferencie de sus competidores (buscadores de Internet, enciclopedias digitales y servicios de suscripciones) en los siguientes aspectos:

- Diferencias respecto de un buscador de Internet: Los documentos de *OCENET Consulta* se han indexado de forma manual utilizando criterios editoriales, lo que hace posible obtener resultados de mejor calidad. Igualmente, la solvencia de los materiales ofrecidos por *OCENET Consulta* es mucho mayor que en una búsqueda en Internet, puesto que los documentos han sido elaborados y contrastados por un equipo editorial formado por expertos en las diversas disciplinas que abarca el portal.

- Diferencias respecto de una enciclopedia digital: En *OCENET Consulta* el número de páginas digitalizadas supera las 300.000 e integra además otro tipo de documentos, como artículos de revista y fuentes primarias, ambos indexados y con idéntica estructura XML. Se puede encontrar en *OCENET Consulta* mayor diversidad de niveles que en una enciclopedia, puesto que contiene desde artículos muy sintéticos a artículos mucho más extensos, a los que se accede mediante un sistema de búsqueda muy avanzado. Por otra parte, *OCENET Consulta* se actualiza con mayor frecuencia que una enciclopedia digital y sus cambios están disponibles de inmediato. Se ha establecido que *OCENET Consulta* actualice y amplíe sus contenidos con una periodicidad de quince días, en general, y con una periodicidad mensual en el caso de los artículos de revistas.

- Diferencias respecto de un servicio de suscripciones: *OCENET Consulta* está en castellano y sus búsquedas siempre ofrecen resultados completos, mientras que mediante suscripción suelen obtenerse únicamente los resúmenes de los artículos. *OCENET Consulta* tiene también la ventaja de que el nivel de los contenidos se ajusta mediante la definición selectiva de los documentos a los que puede tener acceso cada usuario en función de un perfil previamente establecido.

TECNOLOGÍAS LINGÜÍSTICAS EMPLEADAS EN *OCENET CONSULTA*

En el desarrollo de *OCENET Consulta* las tecnologías lingüísticas juegan un papel fundamental en la indexación de los documentos. Esta indexación permite organizar y

clasificar los documentos de la base documental de cualquier manera que se necesite, lo que a su vez hace posible realizar búsquedas muy potentes a partir de las que se obtienen resultados caracterizados por su elevada fiabilidad.

El proceso de creación de contenidos e índices consta de varios pasos. En primer lugar, los textos aportados por Océano se digitalizan y se marcan mediante el lenguaje XML (*Extensible Markup Language*), un procedimiento de codificación de documentos que, en esencia, separa el contenido de la estructura y permite introducir información adicional necesaria para la posterior utilización de un documento. Esta descripción del documento es crucial para que el motor decida qué información ofrecer como respuesta a una consulta.

Una vez marcados, el proceso Delphos desarrollado por CliC/Thera realiza el tratamiento lingüístico –es decir, la normalización y la indexación de los textos-, y los devuelve a Océano para su incorporación a la base de datos documental de *OCENET Consulta*, sobre la que los usuarios finales pueden realizar las búsquedas que deseen a través de la interfaz en la web.

La fase de normalización e indexación es, pues, aquella en la que intervienen las diversas herramientas de tecnología lingüística desarrolladas por CliC/Thera.

Se procede, en primer lugar, a una lematización – extracción de la raíz de las palabras- y a un análisis morfológico –asignación de cada palabra a una categoría como "nombre", "verbo", "pronombre", "adjetivo", etc., con la indicación de género, número, persona, tiempo… - con el fin de hacer posible las búsquedas sin que el usuario deba, por ejemplo, introducir una palabra en singular y en plural. A continuación se procesan los nombres propios para identificar nombres de personas y de lugares, y se regularizan sus formas de modo que, por ejemplo, "Picasso", "Pablo Picasso" y "Pablo Ruiz Picasso" queden señalados como variantes del nombre de la misma persona. También para facilitar las búsquedas se agrupan las palabras sinónimas –como serían "coche" y "automóvil"–, de forma que se puedan localizar documentos que contienen información sobre un mismo tema aunque en ellos aparezcan palabras diferentes a las utilizadas en la consulta.

Los documentos se catalogan automáticamente en función de su contenido, recurriendo para ello a una clasificación por temas previamente establecida y a los descriptores temáticos que se han asignado a cada uno. Finalmente, cuando es necesario se crean vocabularios específicos, propios de una determinada disciplina como la medicina, la geografía, etc.

Una de las ventajas más importantes de la tematización reside en que todas las palabras contenidas en un articulo se remiten al tema del mismo. Ésta es una de las pautas que se combinan en el sofisticado sistema que "decide" la información que se ofrece como resultado a una simple búsqueda del usuario.

Todas estas tareas responden a dos objetivos principales: por una parte, reducir al mínimo el "nivel de ruido" de la consulta, mostrando únicamente los documentos que contienen la información relevante; por otra parte, se pretende así conseguir que los documentos aparezcan en un orden de prioridad adecuado.

CARACTERÍSTICAS DEL SISTEMA

Las ventajas de incorporar tecnologías lingüísticas a *OCENET Consulta* podrían resumirse en los siguientes tres aspectos:

- Se pueden realizar búsquedas más eficaces que en una simple búsqueda de texto completo debido al proceso de normalización, y a la utilización de criterios de relevancia y de significación dentro de la estructura del documento, que hace que los resultados sean más fiables.
- Es factible utilizar los mismos documentos en otros productos con relativamente poco esfuerzo ya que, al estar almacenados en formato XML en una base documental, pueden agruparse e indexarse en "colecciones" independientes. El hecho de que los documentos sean completamente independientes de las colecciones a las que pertenecen permite crear varias estructuras paralelas con el fin de adaptarlos a los requisitos de otros productos.
- Existe la posibilidad de ampliar *OCENET Consulta* con nuevas funcionalidades y niveles de codificación con el fin de dar otros usos a los índices, a los descriptores y a la codificación que se ha introducido, dado que XML permite un trabajo progresivo. Al mismo tiempo, gracias al proceso previo de lematización y de análisis morfológico, se pueden realizar búsquedas por categorías gramaticales, por el uso que se da al lenguaje, por autor, por período o por país, etc.

Trabajar con este tipo de tecnologías implica asumir riesgos inherentes a la aplicación de soluciones innovadoras, fundamentalmente debidos a la falta de códigos de buen funcionamiento establecidos. Podría darse la circunstancia de que tras haber implantado determinadas tecnologías lingüísticas se descubriera que no aportan una mejora sustancial al producto o que tienen como resultado la aparición de información confusa para un público no especializado. Cada vez que se introduce uno de estos procesos y se genera nueva información, es preciso evaluar si ésta va a constituir una ayuda real para el usuario final. Debe también tenerse en cuenta que utilizar tecnologías lingüísticas conlleva un coste adicional cuya rentabilidad no siempre está asegurada de antemano, lo que implica un riesgo que se debe valorar adecuadamente.

CONCLUSIONES

La incorporación de tecnologías lingüísticas a un servicio de consulta de fondos documentales en la web aporta una serie de ventajas en lo que se refiere a la eficacia y la flexibilidad de las búsquedas, así como a la calidad de la información obtenida; permite también la reutilización de documentos en otros productos gracias a la información sistemática y estructurada con la que se han enriquecido los textos electrónicos. El portal desarrollado por *OCENET Consulta* en colaboración con CliC/Thera – *Centre de Llenguatge i Computació* de la Universidad de Barcelona – así como con otras compañías de desarrollo de contenidos y de soluciones informáticas- es un buen ejemplo de los beneficios que pueden obtenerse del empleo de tecnologías lingüísticas, a la vez que

muestra la posibilidad de mejorar el proceso de recuperación y extracción de información si se introduce el conocimiento lingüístico adecuado.

INFORMACIÓN DE CONTACTO SOBRE LAS EMPRESAS:

Editorial Océano:

Dirección postal:	C/ Milanesat, 21-23 08017 Barcelona España
Teléfono:	(+34) 93 280 2020
Fax:	(+34) 93 204 1073
Correo electrónico:	scervell@oceano.com
Persona de contacto:	Sr. D. Sergi Cervell Portillo
URL:	http://www.oceano.com
URL del producto:	http://ocenet.oceano.com

CliC/Thera, Centre de Llenguatge i Computació:

Dirección postal:	Parc Científic de Barcelona Edifici Florensa c/ d'Adolf Florensa s/n 08028 Barcelona
Teléfono:	(+34) 93 403 45 58 / 93 403 56 71
Fax:	(+34) 93 318 98 22 / 93 448 94 34
Correo electrónico:	amarti@fil.ub.es
Persona de contacto:	Dra. Maria Antònia Martí
URL:	http://CliC/Thera.fil.ub.es