

El sistema SARE-Bi de catalogación y recuperación de documentos multilingües

Díaz Labrador, JosuKa*, Abaitua Odriozola, Joseba+, Jacob Taquet, Inés*, Quintana Hernández, Fernando*

*Facultad de Ingeniería, +Facultad de Filosofía y Letras, Universidad de Deusto
Apartado 1 – 48080 BILBAO

[josuka, ines, fquintan]@eside.deusto.es, abaitua@fil.deusto.es

Araolaza, Garikoitz

CodeSyntax

BIC-Berrilan. Azitain Poligonoa P3 E2. 20600 EIBAR

garaolaza@codesyntax.com

SARE-Bi es un sistema de gestión integral de documentos multilingües, que está basado en esquemas de descripción de metadatos que provienen de la anotación de corpora textuales (TEI), de la traducción asistida por ordenador (TMX) y de la localización de software (XLIFF). Todos estos modelos son dialectos de XML que se solapan y complementan de diversas maneras. Dichos estándares se han incorporado y adaptado a un conocido sistema de publicación web (Zope), resultando un entorno cooperativo en que usuarios con diferentes roles (redactores, traductores, administradores) pueden llevar a cabo el ciclo completo de generación y traducción de nuevos documentos, mediante la reutilización de los ya existentes en el sistema.

Palabras clave: Catalogación y recuperación de información, gestión de contenidos multilingües, bases de datos documentales, traducción asistida, TEI, TMX.

1. INTRODUCCIÓN

Una de las consecuencias de la proliferación de información publicada en Internet en formatos y dialectos sucedáneos de HTML ha sido la acumulación caótica de contenidos que dificulta gravemente la gestión y recuperación de información relevante. En los últimos años se han desarrollado propuestas que tratan de paliar este problema.

Una línea importante de investigación ha aplicado sistemas con conocimiento lingüístico que tratan de precisar y acotar, por un lado, el resultado de los buscadores según la acepción más adecuada del término de consulta y de ampliar, por otro, la búsqueda bien a términos semánticamente afines, o bien a textos en otros idiomas [9].

Otra línea ha abordado el problema desde las posibilidades de aplicar a los contenidos publicados en la red la noción de metadato [12, 6]. La apuesta por los metadatos ha ido multiplicando adeptos en los últimos años, sobre todo tras el desarrollo de XML como alternativa a HTML y el espaldarazo que ha supuesto la iniciativa de *web* semántica [1, 14, 4].

En esta comunicación se presenta el sistema SARE-Bi de gestión integral de contenidos

multilingües, que está basado en esquemas de descripción de metadatos que provienen de la anotación de corpora textuales (TEI [10]), de la traducción asistida por ordenador (TMX [7]) y de la localización de software (XLIFF [8]). Todos estos modelos son dialectos de XML que se solapan y complementan de diversas maneras. Dichos estándares se han incorporado y/o adaptado a un conocido sistema de publicación web (Zope [15]).

En las siguientes secciones se recogen, por este orden, la especificación de SARE-Bi y el análisis de situaciones que pretende resolver, la descripción conceptual del sistema, después unas breves ideas de su funcionamiento y su implementación, y finalmente, mejoras que están planteadas para un muy próximo futuro.

2. ESPECIFICACIÓN Y PROPÓSITO

SARE-Bi es un sistema de procesamiento, clasificación y recuperación de documentos multilingües, cuyo propósito es facilitar a los usuarios la realización de las tareas relacionadas con la generación y traducción de nuevos documentos, a través de la reutilización de los ya existentes.

Aunque en principio no parece haber dificultad teórica en aplicar el sistema a documentos de cualquier extensión o ámbito, lo

cierto es que se diseñó pensando primordialmente en documentos de tamaño pequeño o mediano del ámbito administrativo, es decir, documentos internos de una organización con un entorno multilingüe. De hecho, en estos momentos, el sistema SARE-Bi ya está siendo usado en la Universidad de Deusto con un corpus de documentos de distinta índole (avisos, cartas, convocatorias a reunión, normativas, instancias, etc.).

Para ilustrar la funcionalidad del sistema, cojamos como ejemplo las cartas de admisión a la Universidad. El proceso se realiza en tres pasos. Primero el “redactor” compone el documento en una lengua; después los “traductores” generan las versiones en el resto de idiomas, y finalmente el redactor publica el documento multilingüe completo. Se trata normalmente de un proceso repetitivo que se reproduce año tras año. SARE-Bi aprovecha esta circunstancia, proporcionando una alimentación y acceso rápidos a la base documental multilingüe que redactores y traductores van construyendo a lo largo del tiempo.

Las prestaciones de SARE-Bi lo convierten en un complemento perfecto de los gestores de memorias de traducción (Wordfast, Déjà-Vu, etc.), aunque en la práctica acaba absorbiendo muchas de las funciones de estos programas, haciéndolos innecesarios. Su principal ventaja es que ofrece un entorno de trabajo cooperativo en red, en el que tanto el redactor como el traductor comparten la misma base documental, de forma que pueden recuperar los documentos relevantes como punto de partida para la redacción y traducción de versiones actualizadas, sin necesidad de recurrir a ningún otro software de traducción.

Otra situación en la que el sistema resulta muy útil se da cuando el redactor no solo conoce la lengua en la que redacta, sino que conoce también (aunque con menor confianza) la segunda o tercera lengua del documento (situación habitual con el euskera y el inglés, en nuestro caso). Cuando esto sucede, el propio redactor puede usar SARE-Bi para recuperar el documento y practicar las modificaciones en la versión original y en las traducciones. El servicio de traducción se limitará a revisar y validar los cambios introducidos.

Además de agilizar la tarea de traducción, SARE-Bi presenta ventajas como base de datos documental tradicional (monolingüe). Por ejemplo, otro “redactor” puede tener que

escribir una carta de admisión, pero no dispone de ningún ejemplo o plantilla. Con la ayuda del sistema no solo puede recuperar un documento de partida inicial, sino que probablemente pueda disponer de una versión multilingüe.

En resumen, SARE-Bi resuelve el problema tradicional de generación de documentación administrativa, pero, sobre todo, agiliza el proceso de producción multilingüe de dicha documentación. Con ello, se mejora tanto la calidad como la cantidad de documentos multilingües generados, facilitando además la labor de las personas involucradas en las tareas.

3. DESCRIPCIÓN CONCEPTUAL

El sistema SARE-Bi contiene primeramente un *corpus multilingüe anotado, segmentado y alineado*. Es decir, el contenido de los documentos no se almacena tal cual, sino que se aporta un etiquetado que sigue las ideas básicas del estándar TEI. Cada documento se divide en subdocumentos, uno por cada lengua, y el etiquetado aporta primordialmente una segmentación de cada subdocumento, y un alineamiento de los segmentos correspondientes en distintas lenguas. Actualmente, los segmentos son párrafos, y el etiquetado, la segmentación y el alineamiento son procesos que realiza automáticamente el sistema.

En segundo lugar, se asocian a cada documento una serie de *metadatos*, que describen diversos aspectos pragmáticos y que aportan la funcionalidad que se desea para el sistema [3, 13]. El metadato más importante es la *categoría*, que indica la clasificación del documento según una taxonomía jerárquica de distintos niveles (inspirada en otras propuestas de clasificación tipológica [11]), que variará lógicamente con cada implantación del sistema. En el caso particular de la aplicación a la Universidad de Deusto, se diseñó una taxonomía en tres niveles, que, de mayor a menor jerarquía, indican la *función*, el *género* y el *tema* del documento. Por ejemplo, un certificado por asistencia a un cursillo tiene función (primer nivel) “informativa”, es de género (segundo nivel) “certificado”, y su tema (tercer nivel) es “asistencia a un cursillo”. En la actualidad, para la Universidad de Deusto, la taxonomía consta de 3 funciones (“informar”, “inquirir” y “reglamentar”), 25 géneros y 256 temas.

De parecida relevancia es el metadato denominado *estado*, que informa de la situación actual del documento en lo referente a su

multilingüismo. Actualmente, existen tres posibles situaciones para un documento (que siguen un orden jerárquico, de forma que cada una supone un avance respecto a la precedente): *sin_validar* (el texto inicial producido por el redactor), *validado* (versión que cuenta con la aprobación de los traductores) y *normativo* (versión multilingüe correcta que se ofrece como modelo).

También es importante la propiedad de *visibilidad*, que señala el grado de confidencialidad que ha de tener el documento. Dado que el propósito del sistema es almacenar la documentación administrativa de la Universidad, se vió que podrían existir documentos con información sensible, que no deberían hacerse visibles en todas las condiciones. Se han identificado cuatro posibles valores para este metadato: *borrador* (visible solo para el redactor, supone que el documento se encuentra en la fase de elaboración), *confidencial* (visible con fuertes restricciones), *compartido* (visible en la organización, equivalente al concepto de *intranet*) y *público* (visible universalmente, equivalente a información en web).

Los dos metadatos, estado y visibilidad, tienen relación directa con un componente adicional del sistema, como son los usuarios, cuya descripción se recoge más adelante.

Otro metadato relevante es el *centro* (o departamento de la organización) que origina el documento, separado en dos niveles, centro y subcentro. También se almacenan varias *fechas*, concretamente, la fecha original del documento, la fecha de inclusión en el corpus y la fecha de última modificación (aunque esta última se va a desdoblarse en varias como resultado de las mejoras que actualmente se llevan a cabo). Existen finalmente otros metadatos de importancia secundaria para la funcionalidad del sistema.

Ha de decirse que la incorporación de algunos metadatos (los más importantes) no es automática, sino que debe realizarla el usuario cuando añade un nuevo documento al corpus: son la categoría, la visibilidad, el centro y la fecha original del documento. En cuanto al estado, lógicamente, se trata de un metadato cuyo valor irá variando a lo largo del ciclo de edición, y que deberá ser actualizado explícitamente por los usuarios.

Por otro lado, aunque conceptualmente se puede ver el conjunto de documentos existentes en un momento determinado como un corpus

único, en la práctica este puede dividirse en distintos *subcorpus*. Esta división aporta un grado adicional de estructura que puede ser útil en algún momento. En el sistema implantado en la Universidad de Deusto, este aspecto ha tenido gran utilidad, por ejemplo, para pasar sin dificultades de la fase de desarrollo del sistema a la fase final de explotación.

Hay un componente adicional de gran importancia, que es el conjunto de *usuarios*. En principio, se identificaron tipos de usuarios según el *rol* o tarea que realizasen en el sistema, tal como se observó en el análisis de casos recogido en la sección 2. Así, existen *invitados* (usuarios externos a la organización, de “solo-lectura”, con el propósito de que cualquier persona pueda comprobar el comportamiento del sistema), *redactores* (encargados de añadir nuevos documentos, normalmente en una lengua, o multilingües sin revisar), *traductores* (realizan o revisan las traducciones) y *administradores* (gestionan el sistema). Se supone que salvo los invitados (que acceden al sistema por Internet), los demás usuarios son miembros de la organización y que el sistema funciona a este respecto como una *intranet*.

En este primer nivel de descripción, no existe todavía de forma clara un concepto de *permiso* asociado a los usuarios y el metadato estado se encarga de reflejar la situación del documento en cada momento dentro del ciclo de edición.

Sin embargo, el concepto o metadato de visibilidad da origen a la asociación (a los usuarios) de permisos para la realización de tareas en el sistema, apareciendo un metadato adicional, *propietario*, que indica qué usuario introdujo el documento en el sistema. Mediante estos tres atributos, se obtiene finalmente una compleja especificación de permisos según los tipos de usuarios, que se recoge más adelante en la sección 5.

4. FUNCIONES

El sistema SARE-Bi permite básicamente dos operaciones: *inserción* de nuevos documentos y *recuperación* (y visualización) de documentos existentes. También es posible la *modificación* de un documento existente, pero es una variación ligera de la operación de inserción. Por otro lado, la recuperación admite dos modalidades bien diferenciadas: el *filtrado* (o consulta basada en los metadatos) y la *búsqueda de texto* en el contenido de los documentos. Ambas pueden conducir a la

visualización final del documento. Las cinco operaciones resultantes se desarrollan en los apartados siguientes.

4.1. Inserción de documentos

Al añadir un nuevo documento al corpus del sistema SARE-Bi, el usuario aporta en primer lugar los metadatos que no se obtienen automáticamente (categoría, centro, fecha original y visibilidad como más relevantes), teniendo en cuenta que el sistema asigna automáticamente otros metadatos, y en particular, el estado como *sin_validar*.

A continuación, el usuario proporciona el texto de los subdocumentos de cada lengua. El sistema realiza entonces el etiquetado, la segmentación y el alineamiento, y el documento resulta añadido al corpus.

4.2. Modificación de documentos

La modificación del contenido de un documento se puede realizar con una variación sencilla de la función precedente.

Además, también se proporciona (aunque se supone que será utilizada con menor frecuencia) la posibilidad de cambiar algunos metadatos del documento. Sin embargo, precisamente por su carácter primordial al respecto de los permisos asociados a los usuarios (tal como se explicará en la sección 5), los metadatos estado y categoría se presentan aparte.

4.3. Visualización de documentos

Cuando un documento se visualiza después de su recuperación, se muestran los contenidos segmentados y alineados de todas las lenguas de que consta, lo que permite comprobar al usuario que esos procesos se han realizado correctamente. Además, se da una visión no segmentada (o completa) de los contenidos en cada lengua, no solo para facilitar la legibilidad, sino para que sea sencillo, por ejemplo, copiar el texto completo del documento y que este pueda ser utilizado localmente por el usuario (para darle formato, por lo general).

No se muestra información sobre los metadatos, aunque se añade el acceso a la función de modificación. Por contra, se incorporan aquí dos funciones importantes, que son, primero, la exportación al formato TEI de cada subdocumento (por cada lengua), y segundo, la generación de memorias de traducción en el formato TMX para un par de lenguas.

4.4. Recuperación por filtrado

El filtrado de documentos es equivalente conceptualmente a una consulta (*query*) de la base de datos documental, ya que está basado en los posibles valores de los metadatos más relevantes. La figura 1 recoge el formulario que permite establecer los criterios de selección.

Figura 1. Filtrado de documentos

Por medio de filtros

estado: todos

visibilidad: todos

categoría: todos

<.....>

<.....>

centro: todos

corpus: Corpus-2003
TMXtore
XML-Bi
XML-Bi02

ordenar por: actualizado

inverso:

Filtrar

Como se observa, estos criterios hacen referencia a los metadatos estado, visibilidad, categoría (con la taxonomía trinivel apareciendo dinámicamente a través de tres cuadros desplegados), centro y corpus.

Las dos últimas opciones no son criterios de selección, sino de presentación de resultados: pueden ordenarse por centro, categoría, fecha o corpus al que pertenecen, y tanto en orden normal (ascendente) como inverso.

Los resultados se muestran en una tabla, como se ve en la figura 2, con un número relativamente alto de atributos (metadatos) para cada documento, con el propósito de que sea sencillo detectar el que interesa.

Desde esta lista de resultados, el usuario puede acceder a las funciones de visualización y modificación de cada documento, siguiendo los enlaces que se observan a la izquierda y derecha, respectivamente, de la lista.

4.5. Recuperación por búsqueda textual

La otra operación de selección posible es la búsqueda de cadenas de texto en los segmentos de los documentos. La figura 3 muestra el formulario de búsqueda para esta función.

Figura 2. Resultados de un filtrado

Resultados de la búsqueda

Elementos encontrados: 4

N	estado	título	tamaño	lenguas	categoría	centro	corpus	actualizado	fecha doc.	
1	completo	Invitación a conferencia	6	es eu	informar / tarjeta de invitación / acto cultural	ConsejoGob	XML-Bi02	2003/06/13	2001/10/23	Editar
2	borrador	Festival audiovisual	13	es eu	informar / tarjeta de invitación / acto cultural	ConsejoGob	XML-Bi02	2003/06/13	2002/06/29	Editar
3	validado	Decreto de constitución de centro	13	es eu	informar / nombramientos / en general	ConsejoGob	Corpus-2003	2003/06/09	2003/04/08	Editar
4	borrador	Inauguración exposición	7	es eu	informar / tarjeta de invitación / acto cultural	ConsejoGob	TMxtore	2003/06/13	2001/07/28	Editar
actualizar										

Como se observa, la única posibilidad adicional es que se puede especificar la lengua de los documentos a revisar. En cuanto al texto de búsqueda, hay que indicar que en esta primera versión de SARE-Bi las posibilidades son bastante limitadas, pues solo se admiten cadenas literales. Para futuras ampliaciones del sistema, este aspecto constituye una de las mejoras con mayor prioridad.

Figura 3. Búsqueda de texto en segmentos

Búsqueda en texto libre

Texto de búsqueda:

en los idiomas:

Los resultados se muestran en una lista, tal como se observa en la figura 4. Se ofrece en primer lugar el acceso a la visualización del documento. Después, se muestra el segmento que contiene el texto buscado y, si se trata de un documento multilingüe, aparecen los segmentos con que el anterior está alineado (como se ve en el tercer ejemplo de la misma figura). De esta forma, el usuario puede establecer si le interesa el documento encontrado.

Es importante mencionar que aunque no se elija una lengua concreta en el formulario de búsqueda, en el resultado se mostrarán siempre los segmentos de todas las lenguas de que consta el documento.

5. USUARIOS Y PERMISOS

La existencia de usuarios en el sistema no solo permite, como se dijo al final de la sección 3, la separación primordial de tareas o roles, sino también la asignación de permisos, primero, de visibilidad de documentos y, segundo, de realización de las propias tareas. Los metadatos involucrados en la siguiente especificación de permisos son, como ya se ha dicho, el propietario, el estado y la visibilidad.

Los invitados no pueden ser propietarios, lógicamente, y solo tienen permiso para la visualización de los documentos “públicos” (aquellos en que la visibilidad es *público* y el estado *validado* o superior).

Los siguientes usuarios en la jerarquía, los redactores, pueden visualizar todos los documentos salvo aquellos que, teniendo visibilidad *confidencial* o menor, no sean de su propiedad. Por otro lado, pueden añadir documentos (de los cuales quedan como propietarios), y en particular son los responsables de asignarles el valor deseado de

Figura 4. Resultados de una búsqueda

Resultados de la búsqueda en segmentos

Con la búsqueda de **libro**, se han encontrado 3.

1 - [Reclamo sobre fotocopias](#)

es	SÓLO SE PODRÁ REPRODUCIR UN 5% DEL TOTAL DE LAS PÁGINAS DEL LIBRO
013	

2 - [Reclamo sobre fotocopias](#)

es	Título del Libro o Revista:
006	

3 - [Invitación a presentación de libro](#)

es	El Instituto de Derechos Humanos Pedro Arrupe de la Universidad de Deusto le invita a la presentación del libro “El caso Awás Tingni contra Nicaragua: nuevos horizontes para los derechos humanos de los pueblos indígenas” que tendrá lugar el próximo martes 6 de Mayo en la Sala de Conferencias de la Universidad de Deusto a las 7 de la tarde y contará con la presencia de James Anaya, catedrático de Derecho Internacional de la Universidad de Arizona y asesor legal de la comunidad Awás Tigni, y de Mikel Berraondo, investigador del Instituto de Derechos Humanos Pedro Arrupe.
001	
eu	Deustuko Unibertsitateko Pedro Arrupe Giza Eskubideen Institutuak “El caso Awás Tingni contra Nicaragua: nuevos horizontes para los derechos humanos de los pueblos indígenas” liburaren aurkezpena gonbidatzen zaitu. Ekitaldia maiatzaren 6an, asteartean, izango da arratsaldeko 7etan Deustuko Unibertsitateko Hitzaldi Aretoan eta James Anaya, Arizonako Unibertsitateko Nazioarteko Zuzenbideko katedraduna eta Awás Tigni komunitatearen lege aholkularia, eta Mikel Berraondo, Pedro Arrupe Giza Eskubideen Institutuko ikertzailea izango dira bertan.
001	

visibilidad. Sin embargo, solo pueden practicar modificaciones en los documentos de su propiedad que sean *borrador*, e incluso en ellos, no tienen acceso al metadato estado.

Los traductores realizan la tarea de traducción propiamente dicha (aunque también se les permite la inserción de nuevos documentos: en ese caso se aplican los mismos criterios que para los redactores). Se supone que los traductores forman un grupo mucho menos numeroso que el de redactores. Por ello, y dada su labor, se entiende que puedan visualizar y modificar todos los documentos “completos” en cuanto a elaboración (es decir, aquellos que no son *borrador*), incluso los marcados como confidenciales. En particular son los encargados de asignar un nuevo estado al documento a medida que este vaya cumpliendo los hitos que representa dicho metadato, pero no les está permitido modificar la visibilidad u otros metadatos.

Los administradores, finalmente, tienen el conjunto más amplio de permisos, sin restricciones de inserción, visualización, o modificación.

6. CICLO DE VIDA

La forma típica de trabajo en el sistema es el ciclo completo de elaboración de un documento multilingüe. El ciclo empieza cuando un redactor añade un nuevo documento (quizá habiendo utilizado antes las capacidades de filtrado o búsqueda para la reutilización de los ya existentes), del cual queda como propietario, con estado *sin_validar* y visibilidad *borrador*. Cuando el contenido (ya sea en una lengua o en varias) está completo, le asigna el valor deseado de visibilidad (lo más habitual será *compartido*, pero puede ser cualquiera de los tres valores distintos a *borrador*). Tras ello, avisa a un traductor de que el documento está listo.

El traductor accede al documento, y con la función de modificación realiza o revisa la traducción, colocando el metadato estado como *validado*. Entonces, avisa al redactor original de que el documento ya es multilingüe.

Finalmente, el redactor recoge el documento validado, y lo usa localmente (lo habitual será realizar una copia del contenido en el procesador de textos u otra aplicación).

El documento multilingüe permanece a partir de entonces en el sistema. El único caso en que tendría sentido la eliminación de un documento (operación que está reservada a los administradores) sería la acumulación de

documentos similares obtenidos a partir de una plantilla, en que la información lingüística relevante ya está enteramente contenida en el original.

Mediante variaciones de este ciclo típico se pueden resolver las situaciones descritas al principio en la sección 2. Por ejemplo, el traductor puede calificar algún documento como *normativo* (en lugar de simplemente *validado*), por constituir un modelo o plantilla en su categoría, con lo cual un redactor bilingüe puede utilizar dicho documento consciente de su corrección, sin necesidad de la intervención del traductor.

La experiencia acumulada en la recogida del corpus inicial de trabajo nos permite afirmar que esta situación se presenta con frecuencia, sobre todo, en documentos cortos de tipo aviso, cartel o similares, por lo que se espera que el sistema SARE-Bi, al menos en este aspecto, mejore significativamente la situación de bilingüismo de la organización.

7. IMPLEMENTACIÓN

El corpus SARE-Bi está implementado en el sistema Zope [15] como una base de datos orientada a objetos. A pesar de que teóricamente un almacenamiento basado en la tecnología XML parecía lo más lógico, se decidió utilizar Zope, primero, por sus bondades en cuanto al manejo óptimo de la información, segundo, por presentar un complejo sistema de gestión de usuarios con roles y permisos, y tercero, por facilitar (mediante el módulo *Localizer* [5]) la construcción de una interfaz web y multilingüe para el acceso al sistema. De ese modo, este se encuentra plenamente integrado en el sitio web del grupo de investigación. Por otro lado, la funcionalidad XML plena se mantiene gracias a las facilidades para la exportación a los formatos TEI y TMX.

Básicamente, el corpus se modela mediante tres clases de objetos (*DeliTei*, *DeliLang* y *DeliSeg*), que no tienen relación jerárquica (herencia) entre sí, sino de composición (relación “todo/parte” o “tiene”): un *DeliTei* (o documento multilingüe) tiene varios *DeliLang* (subdocumento en una lengua), cada uno de los cuales tiene varios *DeliSeg* (segmentos, en nuestro caso, párrafos, como se sabe). Existe además una clase contenedora, *DeliCorpus*, que puede contener varios documentos multilingües (de la clase *DeliTei*), aportando el grado adicional de estructuración anteriormente

comentado. El diseño de la base de datos orientada a objetos se ha realizado siguiendo las ideas básicas de UML [2].

Tanto los metadatos como el contenido propiamente dicho de los documentos se convierten entonces en *atributos* de los objetos persistentes de la base de datos. Sin embargo, estos atributos (metadatos) no constan como un todo, sino que se pueden separar en conjuntos de atributos independientes (que en Zope se denominan *property sheets*). Además, es posible asignar permisos distintos a cada conjunto de atributos, dependiendo del usuario o rol del mismo. De esta forma, se han podido poner en práctica las especificaciones recogidas anteriormente en la sección 5.

Las funciones de inserción, modificación y visualización, así como las de exportación a TEI y TMX, se conciben como *métodos* encapsulados en las clases de objetos *DeliTei* o *DeliLang*, según el caso. Todo ello conforma lo que se llama un *ZopeProduct*, es decir, una aplicación completa en Zope, que es única independientemente de cuantos sistemas reales se quieran instanciar en un sitio web determinado.

Las operaciones de filtrado y búsqueda, sin embargo, no son parte del *ZopeProduct* estrictamente, sino que se asocian a cada instancia del sistema (la base de datos de objetos donde reside realmente la información), aunque siguen entendiéndose como métodos que se aplican a clases proporcionadas por el propio servidor Zope.

Estas operaciones de filtrado y búsqueda usan el componente de Zope llamado *Catalog*, que realiza internamente una indización de la información, y proporciona una velocidad de respuesta óptima. Básicamente, *Catalog* realiza una indización con varios tipos de índices, pero el sistema SARE-Bi usa únicamente dos: por un lado, los llamados *field indexes*, que se consideran valores atómicos y se utilizan para la operación de filtrado (indizándose los metadatos que tienen sentido en esa operación, como categoría, estado, centro, etc.), y por otro lado, los denominados *text indexes*, en que el dato se trocea en componentes (palabras), usándose para la función de búsqueda textual.

8. LÍNEAS FUTURAS

El sistema SARE-Bi está siendo mejorado en el proyecto X-Flow, en el que se pretende la automatización de los flujos de información que se dan en los distintos procesos asociados.

Una tarea prevista como próxima mejora es automatizar algunos de los procesos que en la actualidad se realizan manualmente, en concreto, la categorización de los documentos. Con este fin, consideramos que va a ser necesario incorporar un método independiente para la definición de taxonomías documentales, que permita mayor portabilidad del sistema hacia otros entornos. El enfoque previsto es que la taxonomía, con los niveles y categorías que se consideren adecuados para cada aplicación concreta, se defina mediante un metalenguaje (basado en XML). El sistema interpretará esta especificación y la integrará para aplicarla en la catalogación de los documentos. Un método similar puede emplearse para la definición de otros metadatos (estados, visibilidad, roles, centros, etc.).

Otro aspecto que se está considerando es la utilización de protocolos que posibiliten la creación de comunidades de sistemas, de forma que distintas instalaciones trabajen de manera cooperativa, a partir de los metadatos, en la elaboración y acceso a catálogos compartidos.

La versión actual del sistema se puede utilizar en la dirección de Internet <http://www.deli.deusto.es/Resources/SareBi> (los detalles para acceder como invitado se dan en <http://www.deli.deusto.es/Resources>).

AGRADECIMIENTOS

El diseño y desarrollo inicial del sistema SARE-Bi han sido posibles gracias a la subvención concedida por el Departamento de Educación, Universidades e Investigación del Gobierno Vasco al proyecto *XML-Bi: Procedimientos para la gestión de flujo documental multilingüe sobre XML/TEI* (PI1999-72), realizado durante los años 2001 y 2002.

Josu Gómez (actualmente en BiText) y Arantza Domínguez se ocuparon, como colaboradores del grupo de investigación, de varios aspectos durante la fase de recogida de datos y diseño.

Luistxo Fernández (CodeSyntax) aportó estimables comentarios en la fase de desarrollo del prototipo.

BIBLIOGRAFÍA

- [1] BERNERS-LEE, Tim (1998): "Semantic Web Road map". Disponible en Internet (18.06.2003) <http://www.w3.org/DesignIssues/Semantic.html>

- [2] BOOCH, Grandy, RUMBAUGH, James y JACOBSON, Ivar (1999): *El lenguaje unificado de modelado*. Addison Wesley.
- [3] CAPLAN, Priscilla (2001): "International Metadata Initiatives: Lessons in Bibliographic Control", en *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium*. Disponible en Internet (18.06.2003) <http://lcweb.loc.gov/catdir/bibcontrol/caplan.html>
- [4] DECKER, Stefan, MELNIK, Sergey, HARMELEN, Frank van, FENSEL, Dieter, KLEIN, Michel C. A., BROEKSTRA, Jeen, ERDMANN, Michael, y HORROCKS, Ian (2000): "The semantic web: The roles of XML and RDF", en *IEEE Internet Computing*, 4(5), pp. 63-74.
- [5] IBÁÑEZ PALOMAR, J. David (2003): "Localizer". Disponible en Internet (18.06.2003) <http://www.j-david.net/software/localizer/>
- [6] KASHYAP, Vinay y SHETH, Amit P. (1998): "Semantic heterogeneity in global information systems: the role of metadata, context and ontologies", en SCHLAGETER, G. y PAPAZOGLU, M. P. (eds.): *Cooperative Information Systems: Current Trends and Directions*, Academic Press, pp. 139-178.
- [7] LOCALIZATION INDUSTRY STANDARDS ASSOCIATION (2003): "Translation Memory eXchange". Disponible en Internet (18.06.2003) <http://www.lisa.org/tmx/>
- [8] OASIS (2003): "OASIS XML Localisation Interchange File Format TC". Disponible en Internet (18.06.2003) <http://www.oasis-open.org/committees/xliff/>
- [9] SPARCK JONES, Karen y WILLETT, Peter, eds. (1997): *Readings in Information Retrieval*, Morgan Kaufman Publishers.
- [10] TEI CONSORTIUM (2003): "Text Encoding Initiative". Disponible en Internet (18.06.2003) <http://www.tei-c.org/>
- [11] TROSBORG, Anna (1997): "Text Typology: Register, Genre and Text Type", en TROSBORG, Anna (ed.): *Text Typology and Translation*, John Benjamins, pp. 3-23.
- [12] WEIBEL, Stuart (1995): "Metadata: The Foundations of Resource Description", en *DLib Magazine*, 1(1). Disponible en Internet (18.06.2003) <http://www.dlib.org/dlib/July95/07weibel.html>
- [13] WITTENBURG, Peter y BROEDER, Daan (2002): "Metadata Overview and the Semantic Web", en ELDA (ed.): *Proceedings of the International Workshop on Resources and Tools in Field Linguistics*, Las Palmas.
- [14] WORLD WIDE WEB CONSORTIUM (2003): "Semantic Web". Disponible en Internet (18.06.2003) <http://www.w3.org/2001/sw/>
- [15] ZOPE COMMUNITY (2003): "Welcome to Zope.org". Disponible en Internet (18.06.2003) <http://www.zope.org/>