[*Translating and the Computer 25*, November 2003 [London: Aslib, 2003]

Metadata for multilingual content management

A practical experience with the SARE-Bi system

JosuKa Díaz-Labrador[1], Joseba Abaitua[2], Inés Jacob[1], Fernando Quintana[1], Garikoitz Araolaza[3]

[1]Facultad de Ingeniería, [2]Facultad de Filosofía y Letras
University of Deusto
Apartado 1. E-48080 Bilbao, Spain
josuka@eside.deusto.es, abaitua@fil.deusto.es,
ines@eside.deusto.es, fquintan@eside.deusto.es

[3]CodeSyntax
BIC-Berrilan. Azitaingo Industrialdea 3K. E-20600 Eibar, Spain
garaolaza@codesyntax.com

Introduction

This paper describes a multilingual document managing system, SARE-Bi, that is based on the use of metadata. In this system, metadata have the role of controlling all phases of a document's life cycle, from the drafting of the first version up to the reutilization of published material, including all intermediate phases of translation, post-edition, validation, publication, and others. The system has been implemented in the web application server Zope and metadata are based on two XML proposals: TEI, for structural mark-up, and XLIFF, for log control. The paper shows how adequate Zope is as an application to manage the entire life cycle of multilingual contents.

Keywords: multilingual document management, metadata, XML, TEI, XLIFF, Zope, machine assisted translation.

SARE-Bi system (http://www.deli.deusto.es/SareBi/)

Problem description and case study

Rapid multilingual delivery of publishable documents is still a challenge for translation technology. Commercial machine translation systems, as for example Systran, Reverso or PAHO's Spanam or Engspan (for the English and Spanish pair), are capable of producing

readable texts, sometimes of unexpected good quality (normally after some period of training or dictionary updating), but which still need rather long and laborious post-editing time before the actual output can be published. Furthermore, very often multilingual publication requires more functions than those usually contained in MT packages.

The problem would normally arise in institutions that are bound to generate, for some reason, documents in two or more languages. This is the case of the University of Deusto which, as any other public institution in the Basque Country, has to publish every official document in at least two languages, Spanish and Basque (*Euskara*), but occasionally also in English or French. What we mean by document is any administrative text that has been made public by some department or centre of the institution. Here we include not only long texts of varying complexity (such as internal statutes, regulations, reports, or proceedings), but also simpler texts including calls, announcements, minutes of meetings, letters, notifications or invitations.

Some fieldwork was carried out in order to find out the actual procedures of both writers and translators. As a result of this, a number of conditioning factors were discovered. In the first place, we found that the production of bilingual or multilingual documents in our institution follows a rather fixed process. One person writes the original document, almost always in Spanish. This is sent to the translation service that generates the versions in the other languages (Basque or English). Then the text is sent back to the original writer, who carries out final editing and takes responsibility for publication.

The original language of most if not all documents is Spanish, because Basque is still a minority language in our community. Although an already large group of people that understand texts in Basque exists, only a few can actually write with sufficient quality in this language. In addition to this, there is a strong tradition, a kind of intertextual inertia, to write in Spanish, particularly in the case of administrative documentation. As a result, the documentation in our organisation is largely written entirely in Spanish, and only later translated into Basque. There is a select but small minority of staff who are capable of producing bilingual documents, although even they on same occasions prefer to use the translation service.

Secondly, we observed that, although a significant quantity of documents that are published in the two languages already exists, the number of documents that are not only written but also published solely in Spanish is very large. Translating has a cost, both in economic terms (many documents are not translated because they are not considered sufficiently important), and of time. Many documents, including smaller-sized ones such as short notifications or calls, have to be generated urgently and very often the translation service has no time to do the job.

In third place, the importance of the documents could be established with regard to their recipients. Documents that address the entire university community (of more than 15,000 people) are among those qualified as very important. More restricted documents, addressing reduced groups of people (a call for a department meeting, for example), are deemed less important. This distinction could lead to the application of "lesser quality"

translation procedure in the case of restricted documents. But what happens normally is that these documents are published only in Spanish. It would not be appropriate to circulate low quality translations.

Finally, one aspect of the fieldwork that attracted our attention was the fact that an increasing number of writers reutilize bilingual published material for some particular types of texts (short letters, calls, announcements or invitations for example). Many short documents like these undergo small changes (of date, place, some parts of the agenda) from one version to the next. We considered this to an interesting starting point. It is normally safe for the editor of the new document to reuse the old version in the text processor, and with little knowledge of Basque to update the changes both in the original Spanish text and in the translation.

In view of this situation our objective was to increase the number of multilingual documents generated in our University, thus reducing both cost in terms of time and money that this effort implies.

How can MT help?

Currently, no system that translates automatically from Spanish into Basque exists, so Machine Translation option cannot be considered. The Basque Government has recently conducted a feasibility study of MT for Basque and, as a consequence, has decided to finance the development of a Spanish/Basque MT system in the near feature. Research groups in the Basque Country (such as IXA, and DELi), as well as two companies from the linguistic-services sector (Elhuyar, Eleka), are expected to take part in the project, although the leading role will be taken by AutomaticTrans, an MT specialised company from Barcelona. In any case, this is a project for the coming years, that will begin at the end of 2003, and which is not presently available.

The only options that can be considered stand in the area of Machine Assisted Translation (MAT). The translation service in our University only partly applies some MAT tools, such as term base. Translation memory systems that have been evaluated and acquired, but have still not been put into operation, due to the time span normally needed before such systems become productive.

Solution (aspect 1): a document management system

Given this situation, we considered that the most practical thing to do was to develop a document-management system with multilingual functionality in which users could find the complete range of document types that are more commonly used within our own institution. This system would allow users to retrieve relevant documents and reuse them to elaborate updated versions. We thought of a cumulative, collaborative system, where different kinds of users like translators or writers, could share their documents, and not only in their definitive form, but also throughout the different stages of elaboration. Such a system would be beneficial not only for its use in the translation process, but also as a document-base fulfilling archiving purposes.

Solution (aspect 2): translation memories

Our group has carried out some basic research in the field of automatic feeding of memory-based MAT systems. In the period 2000-2001, we developed the XTRA-Bi toolkit [Jacob *et al.*, 2001], a set of tools that permits the compilation, segmentation and alignment of bilingual texts. These texts are captured from different Internet sources and then converted into TMX output [LISA, 2003]. The final aim was to combine two complementary technologies: multilingual document management with translation memories.

A system could then be designed in such a way that text would be stored not as a big, and largely blind, repository of translation segments, but as a categorised set of segmented aligned and well-classified parallel documents. Hence this design adds the power of a multilingual document-base to the functionality of a translation memory manager.

Solution (aspect 3): metadata

The last aspect that we considered was the conceptual architecture of the system. Such a system should not be able to manage full documents only, but also document segments. So a broader view of information management was required. We found it in some recent initiatives connected with the evolution of Internet and the shift from rudimentary text mark-up (based in HTML) to solutions derived from the implementation of XML technology.

One of the most negative consequences of the proliferation of information published on the Internet, in various forms and dialects of HTML, has been the chaotic accumulation of contents, which seriously hinders both management and retrieval of relevant information. In recent years several proposals have been made that try to alleviate this problem.

An important line of research has considered the application of linguistic knowledge, firstly trying to make the scope of the search less ambiguous and more precise, and secondly, extending the search either to semantically related terms, or to texts in other languages [Sparck Jones and Willett, 1997]. Another line of research has focused on the notion of *metadata* and its application to content in all its possible uses [Weibel, 1995; Kashyap and Sheth, 1998]. The use of metadata has increased in popularity in recent years, due partly to the development of XML, as a qualified alternative to HTML, and partly to the appealing effects of the Semantic Web initiative [Berners-Lee, 1998; W3C, 2003; Decker *et al.*, 2000].

In this context, for the purposes of text categorisation and cataloguing we have adopted a mark-up solution that is strongly inspired in the guidelines of the Text Encoding Initiative (TEI), a well-known standard in the field of corpus linguistics [TEI Consortium, 2003; McEnery and Wilson, 1996]. The emphasis in our use of TEI is not so much on the linguistic aspects of the texts, but on basic structural aspects and on the set of metadata that covers cataloguing information on the TEI header.

In sum, SARE-Bi can be defined as a multilingual document management system that allows collaborative and incremental compilation of documents, that uses metadata as a conceptual mechanism for controlling all aspects of the document base and which shows a strong resemblance to memory-based machine translation systems.

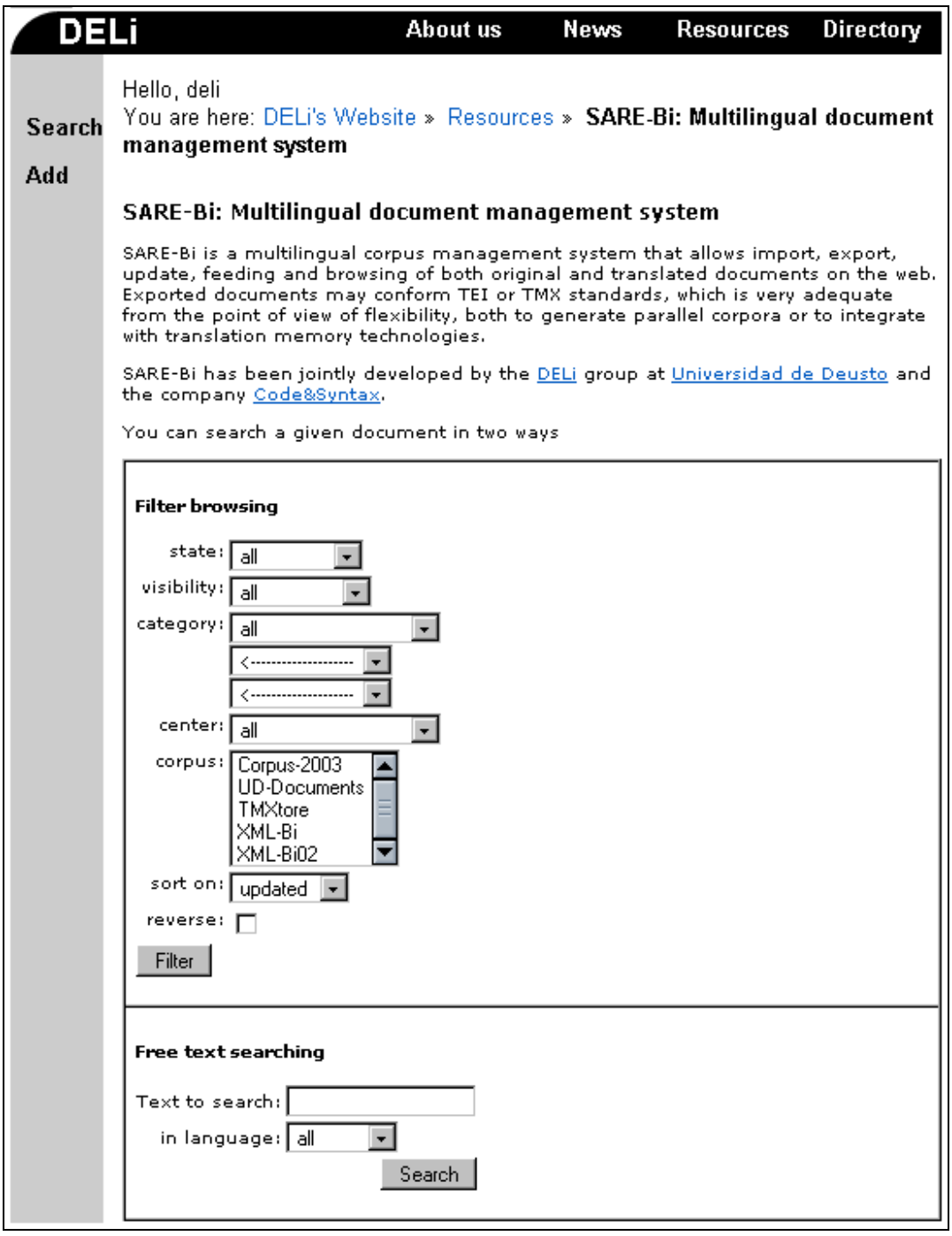


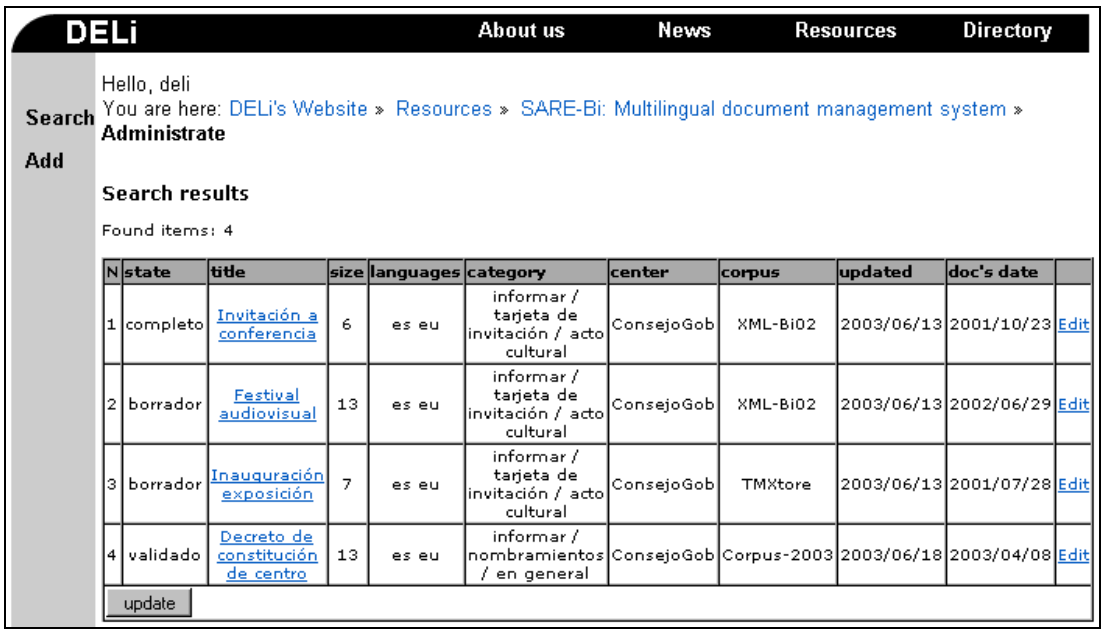
Figure 1: SARE-Bi's front page

A first tour on SARE-Bi

We will start by illustrating the most salient functionality features of SARE-Bi. The system's front page (http://www.deli.deusto.es/SareBi/) is as Figure 1.

As can be seen on the menu on the left part of the screenshot, there are two main operations a user could do in SARE-Bi: document *search*, and *adding* a new document. Let us concentrate on the first operation, which is indeed the most frequent one, and because of that, it is implemented on the main entry to the system.

There are two ways of performing a search: a *filter browsing*, and a *free text search*. In the first mode, the user is allowed to retrieve (filter) a list of documents that meet certain criteria, precisely those which have been associated to them through the use of metadata. Using the first form of the main page to perform a filtering, as we will see later, the user could obtain a list of documents like those shown in figure 2.



N	state	title	size	languages	category	center	corpus	updated	doc's date	
1	completo	Invitación a conferencia	6	es eu	informar / tarjeta de invitación / acto cultural	ConsejoGob	XML-Bi02	2003/06/13	2001/10/23	Edit
2	borrador	Festival audiovisual	13	es eu	informar / tarjeta de invitación / acto cultural	ConsejoGob	XML-Bi02	2003/06/13	2002/06/29	Edit
3	borrador	Inauguración exposición	7	es eu	informar / tarjeta de invitación / acto cultural	ConsejoGob	TMXtore	2003/06/13	2001/07/28	Edit
4	validado	Decreto de constitución de centro	13	es eu	informar / nombramientos / en general	ConsejoGob	Corpus-2003	2003/06/18	2003/04/08	Edit

Figure 2: Results of a filtering

The user can see the title of the document in each row of the table, along with the values of several metadata associated to it. If the user clicks on the title of a document, then he could visualise its contents, as we can see in figure 3.

The user could read the entire document (without any separation, as a whole) in the first part of the visualisation screen. Alternatively, she could look at the document in a segmented, aligned form, below in the same page. In this way, the multilingual correspondence of the parts of the document is made explicit.

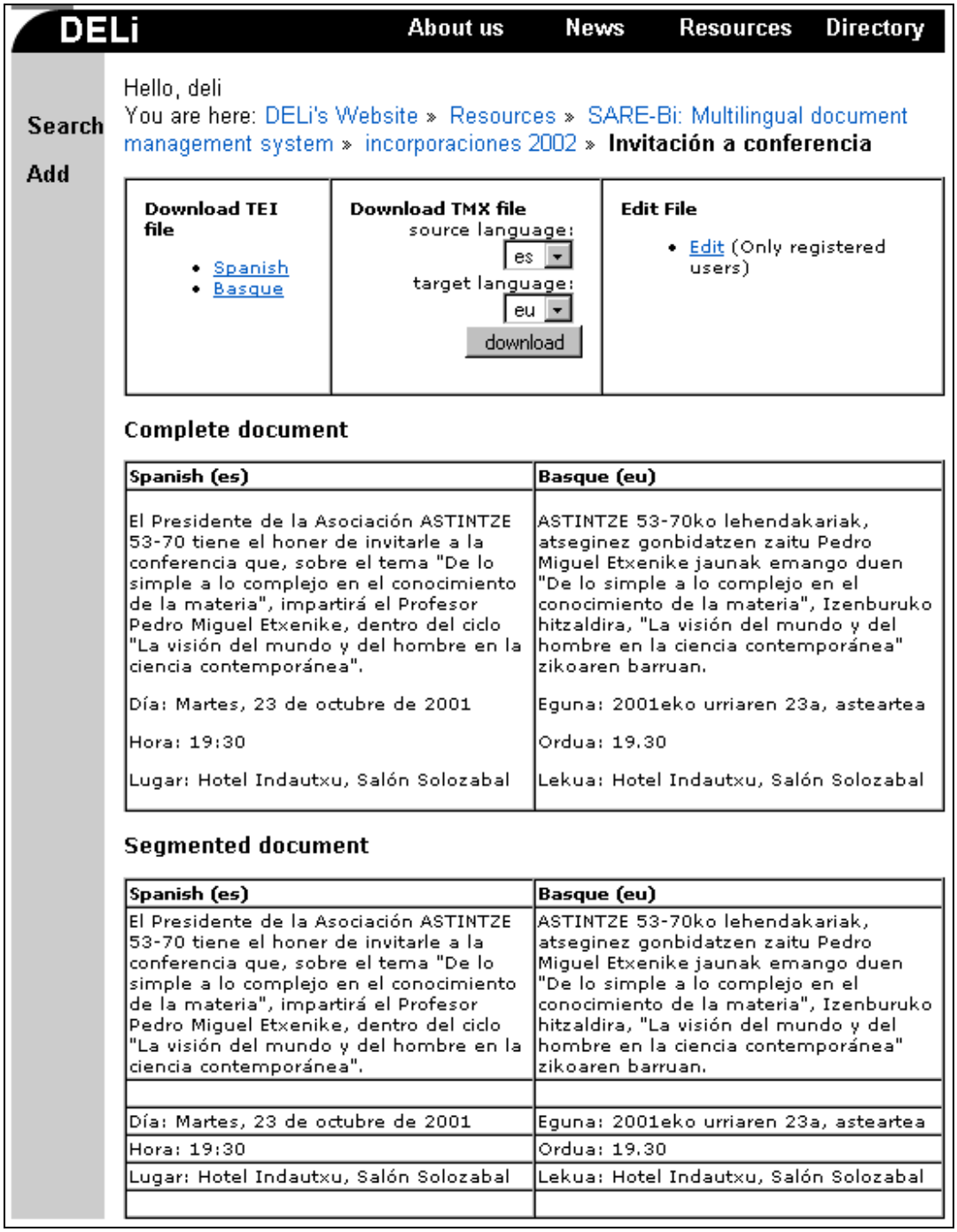


Figure 3: Viewing document contents

Finally, at the very top of the document page, the user is able to export the contents to the TEI and TMX formats. In the first case, only a monolingual version is generated. In the second, a bilingual set of translation memories is obtained, which of course are immediately ready to feed any MT software using that technology.

The other kind of search that a user could perform is a *free text searching*, using the second form that appears in the main page (see figure 1). He could write a word in the textbox to obtain a list of segments (parts of the documents) that contain that word. An example result is shown in figure 4.

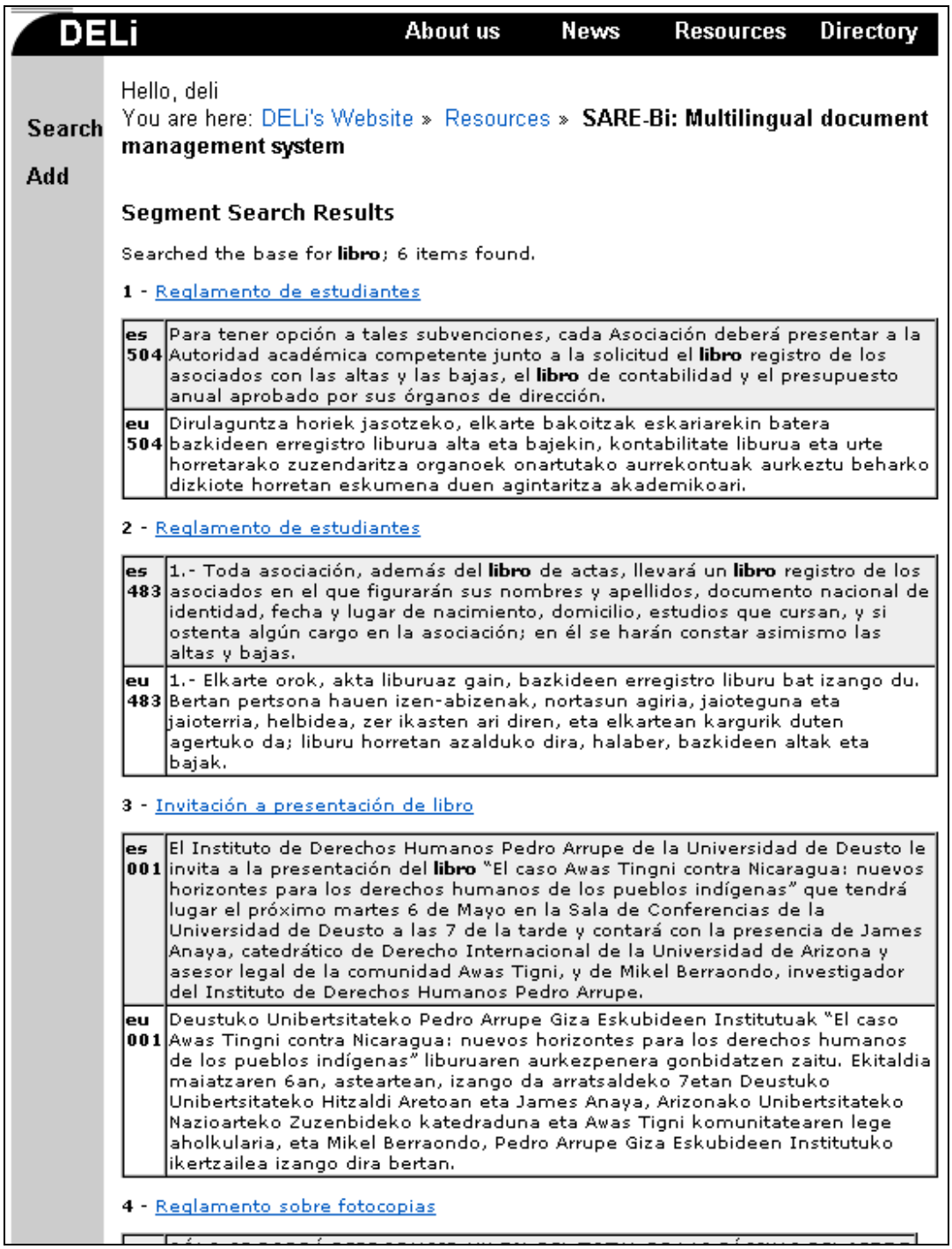


Figure 4: Results of a free text search

The result of this search shows the segments that contain the required word. Actually, not only is the segment of the language of the word shown, but the rest of corresponding segments (in the other languages of the document) as well. In this way, the system acts like a translation memories browser, giving the user the possibility of knowing the translation of some words or expressions in other languages.

Following the links associated to the title of each document shown, the user could reach the whole document visualisation page, in the same way as figure 3 shows.

The other main operation that a user could perform in SARE-Bi is adding a new document. This is a twostep function: in the first, the user gives the values of non-automatic metadata that will be associated to the document, and creates as many (empty) subdocuments as it has languages (figure 5).

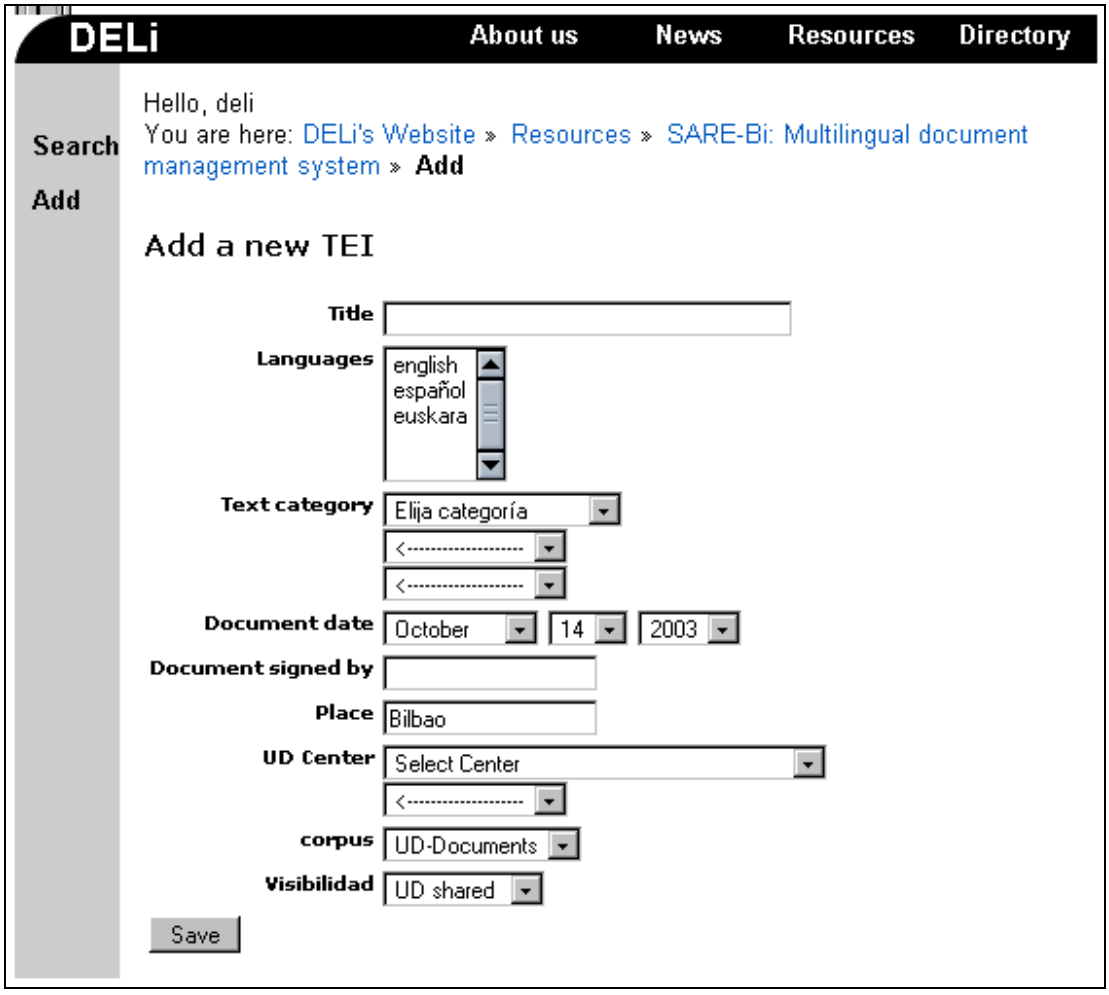


Figure 5: Adding a document, first step

In the second step, the user provides the contents of each subdocument (figure 6). Mark-up, segmentation, and alignment of segments are done automatically by the system.

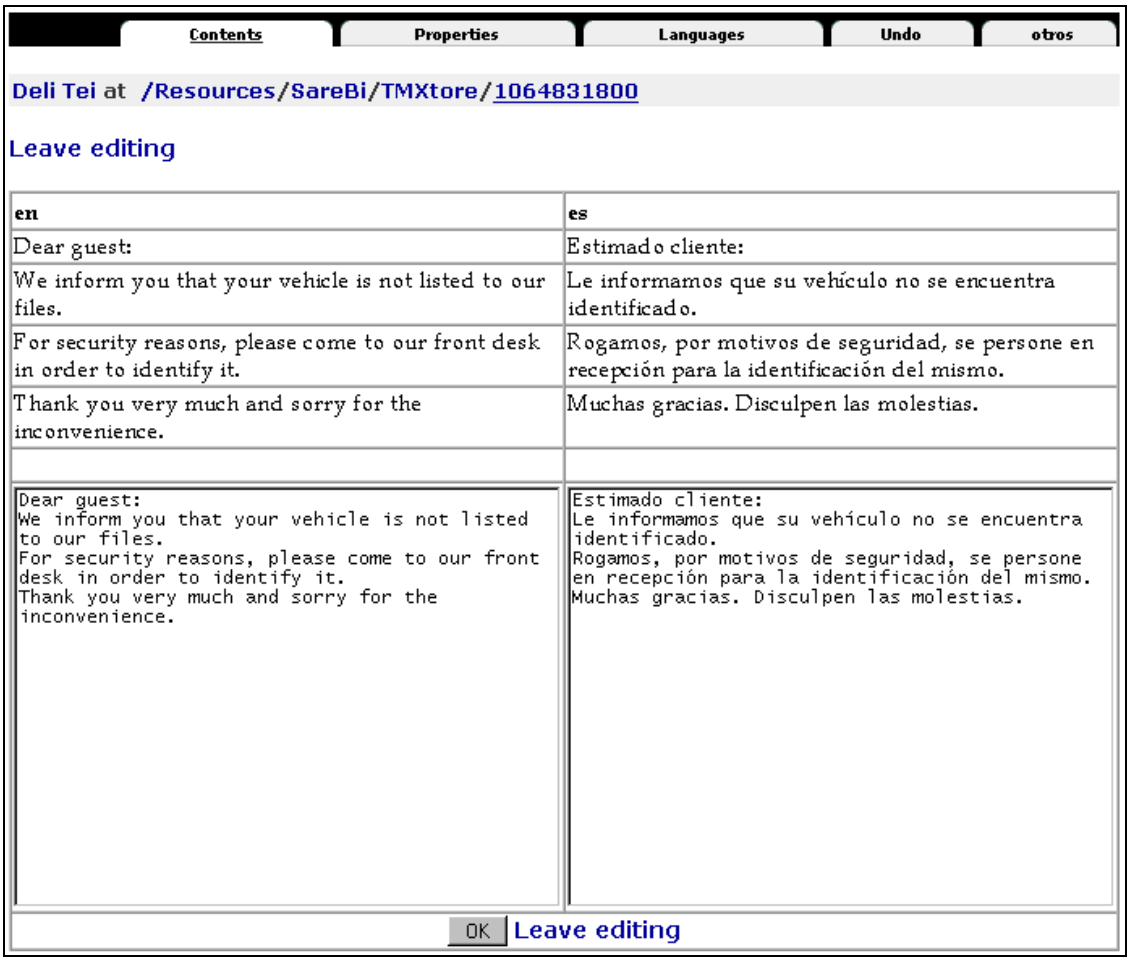


Figure 6: Adding a document, second step

It is important to mention that this operation may be used to add a monolingual document, and later the same user (or indeed, another) could add the subdocument in another language. In fact, the typical document generation cycle will involve the creation of a monolingual document by a user, and the later addition of a translated subdocument by a translator, for example. To this end, the screenshot of figure 6 is the common modification page for a given document: once created, any user with enough modification permissions can reach this page from the links labelled “Edit” on the pages shown in figures 2 and 3.

Conceptual description of SARE-Bi

SARE-Bi system contains, first of all, a multilingual annotated, segmented, and aligned *corpus of documents*. The contents of each document are firstly annotated (in a TEI-like fashion). Fundamentally, this annotation gives a segmentation of the subdocument in

each language, and an alignment of the corresponding segments in different languages. At the moment, segments are paragraphs, and annotation, segmentation, and alignment are automatically carried out by the system. On the other hand, although conceptually the existing document set can be seen as a whole corpus, it is actually divided into different sub-corpora. This division brings an additional degree of structure that users may find helpful.

Secondly, a series of *metadata* is associated to each document, which describe its diverse pragmatic features and which contribute to the functionality that is desired for the system [Caplan, 2001; Wittenburg and Broeder 2002]. The most important metadatum is the one that classifies the document, according to a hierarchic taxonomy of different levels. In the application to the University of Deusto, this taxonomy has three levels that indicate the *function*, the *genre*, and the *topic* of the document (this is inspired in known typological classification proposals [Trosborg, 1997]). For example, a certificate of attendance at a short course is classified at the three levels: the first is at the function level (*informative*), the second is genre (a *certificate*) and the third is the topic (the *attendance* at the course). At the present time, the University of Deusto taxonomy consists of 3 functions, 25 genres and 256 topics. In figure 7, we show the first two levels of the hierarchy, whereas figure 8 contains an excerpt from the full three-level taxonomy.

10000/reglamentar	
11000/	autorización
11100/	acuerdo
11200/	instrucciones
11300/	normativa
11400/	bases
11500/	plan
11600/	ceremonial
20000/informar	
21100/	aviso
21200/	carta
21300/	saluda
21400/	certificado (por)
21500/	convocatoria
21600/	tarjeta de invitación
21700/	folleto (imprenta y web)
21800/	guía
21900/	memoria
22000/	catálogo
23000/	actas
23100/	anuncios en prensa
23200/	carteles de propaganda
23700/	nombramientos
30000/inquirir	
31100/	ficha
31200/	impreso
31300/	cuestionario
31400/	instancia

Figure 7: First two levels of SARE-Bi hierarchical document taxonomy

30000/inquirir
31100/ ficha
31101/ aceptación o renuncia de beca
31102/ boletín de inscripción
31103/ datos de viaje
31104/ modelo de pago
31105/ relación de coordinadores departamentales
31106/ planificación actividad de profesores
31107/ prácticas
31108/ datos estadísticos
31109/ boletín subscripción revista
31200/ impreso
31201/ de solicitud de beca
31202/ de solicitud de expediente
31203/ de solicitud de admisión
31204/ de solicitud de alojamiento
31205/ de programa Sócrates
31206/ de matrícula
31207/ factura
31208/ recibí
31209/ petición de fotocopias
31210/ permiso
31211/ permiso para asistencia a congreso
31212/ justificante de examen
31213/ solicitud proyecto de investigación en líneas priorizadas
31214/ reclamación de beca
31215/ solicitud beca BBK
31216/ solicitud beca Sasakawa
31217/ solicitud página web
31300/ cuestionario
31301/ comunicación
31302/ evaluación de asignatura
31303/ evaluación de curso
31304/ perfil académico profesional
31305/ evaluación docencia
31306/ autoevaluación docencia
31307/ evaluación cursos drogodependencias
31308/ evaluación Euskal Irakaslegoa
31309/ satisfacción padres alumnos ESIDE
31310/ satisfacción empresas con titulados ESIDE
31400/ instancia
31401/ inscripción pruebas mayores 25 años
31402/ solicitud de adaptacón de planes de estudio
31403/ solicitud de convalidación asignaturas
31404/ solicitud de reconocimiento complementos
31405/ solicitud de reconsideración admisión
31406/ solicitud de título
31407/ solicitud de traslado expediente
31408/ solicitud cambio de asignaturas optativas y libre elección

Figure 8: Excerpt from the full three-level hierarchical document taxonomy of SARE-Bi

Metadata that inform of the *state* and the *visibility* (confidentiality) of the document are also very important, and special indeed. Metadata other than these are *static*: they are assigned at the moment of creation of a document, and usually, they do not change (unless there were any mistakes). On the contrary, state and visibility are *dynamic*: their values change throughout the edition cycle to show the composition/multilinguism situation of the document. To this end, users are given one of the tabs shown in the document modification page (figure 6), as seen in figure 9.

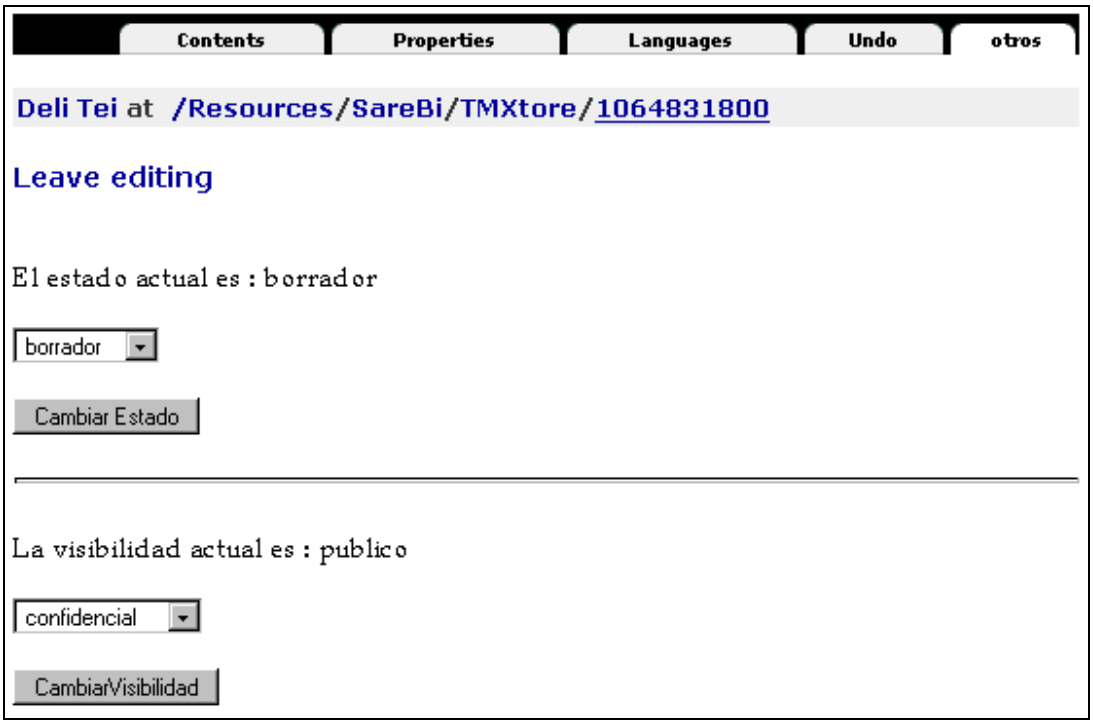


Figure 9: Modifying state and visibility metadata

The state indicates the stage of translation of the document: it may be *non-validated* (in the course of translation, or monolingual), *validated* (already translated, or accepted as a valid translation), and *normative* (multilingual version of special relevance, that is offered as a model). The visibility assigns the degree of confidentiality of the document: when it is at the elaboration stage, it is a *rough draft* (visible only to its owner), whereas when it is already finished (at least in a monolingual version) it may be *confidential* (visible with a lot of restrictions, for documents with sensitive information), *shared* (visible to all users of the system in the organisation, equivalent to the concept of intranet), and *public* (visible for any user connected to the system from the web). These two metadata, state and visibility, are directly related to another important component of the system, that is the set of users, which is considered later in this section.

Another important metadata is the centre (or department of the organisation) that originates the document, separated into two levels, *centre* and *subcentre*. In addition, several *dates* are stored, that is to say, the original date of the document, the date of inclusion in the corpus, and the date of the last modification.

It is important to point out that the assignment of some metadata (the most relevant - the category, the centre, and the original date of the document) is still not automatic: the user has to assign them when he adds a new document to the corpus. On the other hand, state and visibility have a dynamic behaviour, controlled by the users, throughout the edition cycle, because they are users who should change their values for a document as it reaches the different stages of composition and translation.

There is an additional component of the system that is the *set of users*. In the first stages of the design of the system, kinds of users were associated to the different tasks allowed in it. From that perspective, there are four types of users:

1. *Guests* are users outside the organisation; they can interact with the system in a "read-only" way, as a demo.
2. *Writers* are the people who develop new documents.
3. *Translators* are, obviously, the users from the translation bureau.
4. *Administrators* are those who maintain the system.

We can suppose that, apart from the guests (who can access to the system universally from Internet), the rest of the users are members of the organisation, and that the system works for them more like an intranet.

Later on, in the development of the system, we saw the need for defining *permissions* associated to the tasks performed by users. Then, there is an additional metadata of a document, the *owner*, which is the user who first created the document. Depending on the owner, and the state and visibility metadata of the documents, we can define a complex set of permissions for the tasks allowed to users:

1. *Guests* cannot be owners (they cannot add new documents nor modify the existing ones), and they only have the right to visualise the so-called "public" documents: those with "public" visibility and state at least "validated".
2. *Writers* can visualise the contents of any document except those with "confidential" or "draft" visibility of another owner. They can add new documents, and remain owners of them; in particular, each user of this kind is responsible for assigning the correct value of the visibility metadatum of their own documents. However, they can only modify documents of their own, and in fact, they cannot access state metadatum.
3. *Translators* have a larger set of permissions, because of their task. They can, of course, also add new documents, as if they were writers, under the same conditions that apply to this kind of user. They then have the right to visualise and modify the contents of any document except those in elaboration (visibility "draft"); they have specific access to the "confidential" documents. Translators are responsible for setting the correct values of the state metadatum, as the document undergoes the different stages of translation. However, they cannot modify either the visibility or other metadata.
4. *Administrators* have no restrictions of any kind.

Typical edition cycle

Let us complete the description of tasks previously given on the first tour on SARE-Bi with a full view of the typical document generation cycle allowed by the system.

To get a clearer idea, let us take the example of the admission letters that are sent to course applicants. The secretary of the centre (a user of the "writer" kind) first performs a filter browsing or a text search to find out if there is any letter of this kind already in the system. Suppose that none is found. Then he creates a new document, only in Spanish, with the required contents. On creation, visibility is "draft" and state is "non-validated" by default. When he finishes content introduction, he must assign the definitive degree of visibility to the document - normally "shared".

Then, he calls the translating bureau, asking them to translate the document to Basque. A user type "translator" does the translation, adding the language Basque and its contents to the document. When the translation is done, the translator assigns the final value of the state metadatum, normally "validated", and calls the secretary back. After that the multilingual document can be retrieved.

But suppose that after performing the filtering, a previous admission letter is found. Accessing its contents, the secretary could reuse it to compose the new letter. Suppose now that the secretary knows Basque well: as the document is not very complex, he could made the necessary changes in the two languages to obtain a bilingual document, without any translation bureau intervention. On the contrary, suppose the secretary knows Basque a little. He could try the changes both in Spanish and in Basque, but as he is not confident of the result, he calls the translator. Note that normally in this case, there will be minor errors in the document, so the translator's work will be considerably easier.

A translator could also assign the "normative" state to a given multilingual document, when it is paradigmatic in some way: it may contain a special important terminology, it may be a template in its category. The normative documents are then those that could be used by bilingual writers with a high degree of confidence.

Implementation of SARE-Bi

SARE-Bi has been implemented in the Zope system [Zope Community, 2003] under the conceptual scheme of object-oriented databases. Although theoretically a storage based on XML technology seemed appropriate, it was decided to use Zope due to its optimal handling of information, its support for the basic searching and retrieval operations, and its facilities for multilingual web interface construction (by means of the *Localizer* module [Ibáñez Palomar, 2003]). That way, the system is totally integrated into the web site of the research group, and it may be used collaboratively by any number of users. On the other hand, total XML functionality is achieved by means of supported export operations to the TEI and TMX formats.

The design of the object-oriented database has followed the ideas of UML (Unified Modeling Language [OMG, 2003]). The class diagram is shown in figure 10. Basically, a

set of documents, or corpus, is modelled with three classes of objects (*DeliTei*, *DeliLang* y *DeliSeg*), which have no hierarchical relation (i.e., inheritance), but of composite aggregation one (i.e., a "whole/part" relationship):

1. A *DeliTei* (multilingual document) has several *DeliLang* (one-language subdocument).
2. A *DeliLang* has several *DeliSeg* (segments, which in our implementation are paragraphs).
3. A *DeliSeg* contains the textual contents of a paragraph.

A container supra-class, named *DeliCorpus* also exists, which can contain as many documents (*DeliTei*) as required. Then, the database is defined as a set of persistent objects of the *DeliCorpus* class. These four classes inherit from classes provided by Zope, as *ZObject*, *CatalogAwareBase*, and *ZObjectManager* to achieve part of their functionality.

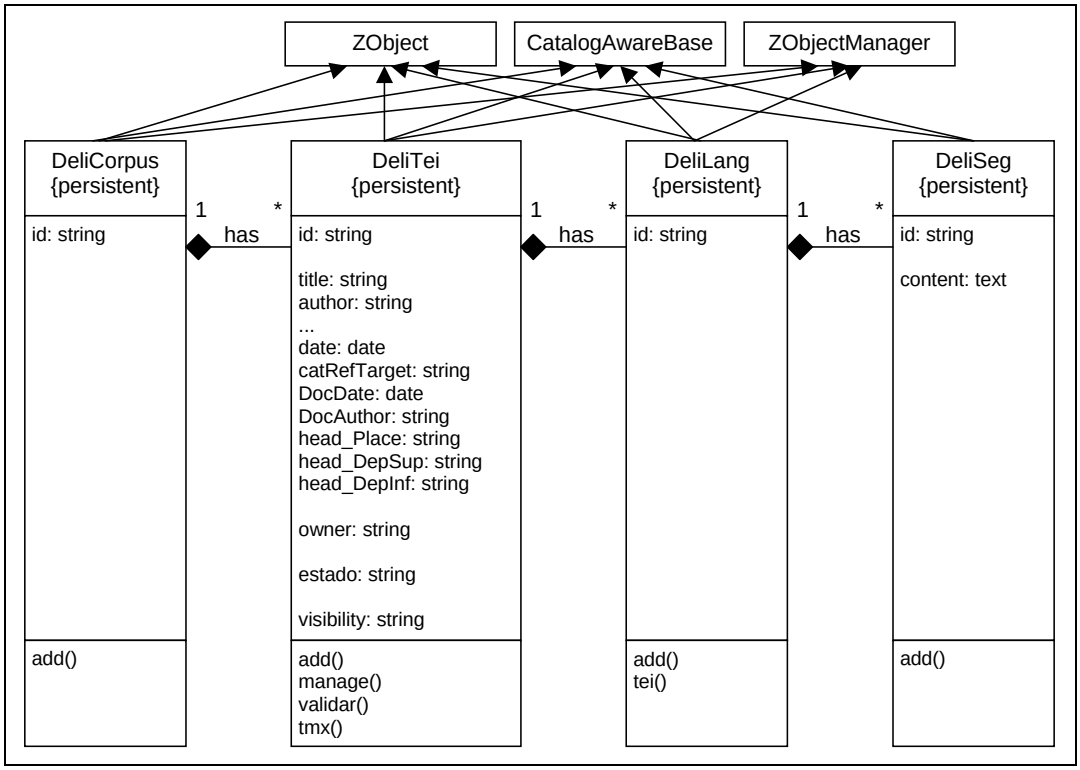


Figure 10: SARE-Bi class model

From this perspective, both the metadata and the proper contents of the documents become *attributes* of the database persistent objects. As figure 10 shows, the textual content of each document is the only attribute of the *DeliSeg* objects. On the other hand, all the metadata are attributes of the *DeliTei* class, due to the fact that metadata are associated to entire documents. Nonetheless, Zope allows for the arrangement of several attributes into disjoint sets, the so-called *property sheets*, which can be handled

independently. This feature has been crucial, for example, for the implementation of the user management policies stated previously.

Both the low-level functions related to document management (insertion, change, visualisation, metadata modification) and the export facilities (to the TEI and TMX formats) are encapsulated methods of the classes *DeliTei* and *DeliLang*. All of this constitutes a so-called *Zope Product*, that is, the basic engine of the system, which is ready to be used in any required application.

However, neither the filtering and searching operations, nor the final web interface are a part of the basic system, but an instance or application of it. That is, they are, to a large extent, independent of the Zope Product. Although this is more a consequence of the Zope internal object-oriented architecture, it does have advantages because, for instance, it allows us to quickly develop the application for another organisation: only a few minor customisation changes on the basic engine (the Zope Product) are required, and effort may be focused on the interface.

Filtering and searching use the *Catalog* Zope component. It makes an internal indexing of the required data and gives a very good query response. *Catalog* allows for several types of indexes to be defined, although SARE-Bi system uses only two of them:

1. The so-called *Field Indexes* refer to atomic data (i.e., values considered as a whole). They are used to perform document filtering, for which several metadata (category, state, visibility, centre, and so on) are indexed.
2. The *Text Indexes* break text up into individual words, so they are suitable for the free text searching function.

The web interface is a multilingual one, in Spanish, Basque, and English (although there are still some items pending translation). It is based on the facilities provided by the Zope module *Localizer*, that allows for easy content localisation.

Conclusions

SARE-Bi is the main result of the XML-Bi project [Abaitua *et al.*, 2001]. The system has been used as a finished application since May 2003. The actual group of users is composed of six writers (from different centres in our University) and two translators from the translation service (who, as a matter of fact, had been working with the system several months earlier). This group is an experimental one, and in a few months, depending on the evaluation of the users, it will be decided whether or not our organisation will adopt the system as a working standard.

So, it is too soon to present any measuring of goal fulfilment. For the moment, we can find a sustained increment in the number of documents, and we have the (mostly) positive comments from the users, so we expect SARE-Bi to achieve its goals.

Meanwhile we are working to improve the system. The X-Flow (“Multilingual content workflow management with XLIFF and TMX”) project [Jacob *et al.*, 2003] is aimed at

the automation of the workflow tasks that users must perform in SARE-Bi. For instance, at the moment, a writer has to call the translator to tell him that a new document has been completed, and the translator has to make the call back when the translation is finished. Our plan is to mechanise these tasks. When a user (either writer or translator) logs in the system, several alerts will inform him of the document workflow: a writer will receive an alert when a document gets translated, a translator will be informed that there are documents waiting for translation, and so on.

On the other hand, X-Flow will include full management of document versioning, to capture all the stages of document edition cycle. To this end, we are applying the new XLIFF (*XML Localisation Interchange File Format*) standard [OASIS Open, 2003], which is "a format to store extracted text and carry the data from one step to another in the localisation process" [opentag.com, 2003].

Another line of improvement in SARE-Bi is related to the application of emerging linguistic engineering technologies. It must be said in advance that, from the very beginning of the design of SARE-Bi, we consciously tried to avoid sophisticated avant-garde procedures, such as automatic categorisation tools, language recognisers, lemmatisers, segmenters, aligners, and so on, but to rely on efficient and well-tested tools. Our goal was above all to develop the minimal, not-a-toy production system that could fulfil the basic document management needs of the members of a given organisation.

Well, we already do have this, so it is now time to add some of the mentioned tools. We would have to analyse the advantage of adding each new procedure, because we believe that, in some cases, an automation tool could not produce the supposed benefits. For instance, one could think of a finer granularity for the segmentation of contents, recognising not only paragraphs, but also sentences, phrases, terms, proper names, acronyms, and so on. It is not apparent, at first sight, that this improvement in the linguistic information mark-up would mean an improvement in the use of the system. So we plan to make an in-depth analysis of the actual advantages that each automation tool would have on the system before going on to its development and implementation.

References

Abaitua, J.; Domínguez, A.; Isasi, C.; Ramírez, J.L.; Jacob, I.; Madariaga, I.; Casillas, A.; Martínez, R.; Garay, A. and Diedrich, T. [2001] "XML-Bi: procedimientos para la gestión de flujo documental multilingüe sonre XML/TEI", *Procesamiento del Lenguaje Natural*, 27 (september), 293-294.

Berners-Lee, Tim [1998] "Semantic Web Road map", http://www.w3.org/DesignIssues/Semantic.html.

Caplan, Priscilla [2001] "International Metadata Initiatives: Lessons in Bibliographic Control", *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium*, http://lcweb.loc.gov/catdir/bibcontrol/caplan.html.

Decker, Stefan; Melnik, Sergey; Harmelen, Frank van; Fensel, Dieter; Klein, Michel C. A.; Broekstra, Jeen; Erdmann, Michael and Horrocks, Ian [2000] “The semantic web: The roles of XML and RDF”, *IEEE Internet Computing*, 4(5), 63-74.

Ibáñez Palomar, J. David [2003] “Localizer”, http://www.j-david.net/software/localizer/.

Jacob, I.; Abaitua, J.; Díaz, J.; Gómez, J. and Ocina, K. [2001] “XTRA-Bi: extracción automática de entidades bitextuales para software de traducción asistida”, *Procesamiento del Lenguaje Natural*, 27 (september), 305-306.

Jacob, I.; Abaitua, J.; Díaz, J. and Quintana, F. [2003] “X-Flow: gestión de flujo de contenidos multilingües sobre XLIFF y TMX”, *Procesamiento del Lenguaje Natural*, 31 (september), 317-318.

Kashyap, Vinay; and Sheth, Amit P. [1998] “Semantic heterogeneity in global information systems: the role of metadata, context and ontologies”, in Schlageter, G. and Papazoglou, M. P. (eds.) *Cooperative Information Systems: Current Trends and Directions*, Academic Press, 139-178.

LISA (Localization Industry Standards Association) [2003] “Translation Memory eXchange”, http://www.lisa.org/tmx/.

McEnery, Tony and Wilson, Andrew [1996] *Corpus Linguistics*. Edinburgh University Press.

OASIS Open [2003] “XLIFF 1.1 Specification, Committee Specification, 25 Jun 2003”, http://www.oasis-open.org/committees/xliff/documents/xliff-specification.htm.

OMG (Object Management Group) [2003] “OMG Unified Modeling Language Specification, March 2003, Version 1.5, formal/03-03-01”, http://www.omg.org/technology/documents/formal/uml.htm.

opentag.com [2003] *XLIFF*, http://www.opentag.com/xliff.htm.

Sparck Jones, Karen and Willett, Peter (eds.) [1997] *Readings in Information Retrieval*, Morgan Kaufman Publishers.

TEI Consortium [2003] “Text Encoding Initiative”, http://www.tei-c.org/.

Trosborg, Anna [1997]: “Text Typology: Register, Genre and Text Type”, in Trosborg, Anna (ed.) *Text Typology and Translation*, John Benjamins, 3-23.

Weibel, Stuart [1995] “Metadata: The Foundations of Resource Description”, in *DLib Magazine*, 1(1), http://www.dlib.org/dlib/July95/07weibel.html.

Wittenburg, Peter and Broeder, Daan [2002] “Metadata Overview and the Semantic Web”, in ELDA (ed.) *Proceedings of the International Workshop on Resources and Tools in Field Linguistics*, Las Palmas, Spain.

W3C (World Wide Web Consortium) [2003] “Semantic Web”, http://www.w3.org/2001/sw/.

Zope Community [2003] “Welcomo to Zope.org”, http://www.zope.org/.

Acknowledgements

SARE-Bi system development has been partly funded by the Autonomous Basque Government, Department of Industry (project X-Flow, 2002-2003) and Department of Education, Universities, and Research (project XML-Bi, 2000-2001).