

Categorización textual y flujo documental en la gestión de contenidos multilingües para entornos web

Joseba Abaitua, Inés Jacob, JosuKa Díaz, Fernando Quintana

Grupo **DELi**



/

- Introducción
- SARE-Bi
- Lineas futuras

- El grupo DELi
- El problema
- Nuestra solución

Los comienzos... 1998,

- profesores de las facultades de Letras y ESIDE,
- con intereses comunes en las áreas de edición digital e ingeniería lingüística.

Áreas de investigación:

- Humanismo para la sociedad
- Tecnologías de la Información y Comunicación



- El grupo DELi
- El problema
- Nuestra solución

Gestión eficaz

- de contenidos que se publican en más de un idioma,
(por ej. portales web multilingües de grandes instituciones y empresas)

Control global del flujo

- de versiones de textos desde redacción hasta publicación

Un entorno real: Universidad de Deusto

- documentos internos de una organización multilingüe
(por ej. avisos, cartas, convocatorias a reunión, normativas, instancias, etc. . .)

- ✓ ● El grupo DELi
- ✓ ● El problema
- Nuestra solución

SARE-Bi

- sistema de procesamiento, clasificación y recuperación de documentos multilingües

Su propósito general es

- facilitar la tareas de generación y traducción de nuevos documentos
- a través de la reutilización de los ya existentes

- ✓ ● El grupo DELi
- ✓ ● El problema
- ✓ ● Nuestra solución

/

- ✓
 - Introducción
 - SARE-Bi
 - Lineas futuras

- Casos de uso
- Descripción conceptual
- Funcionamiento
- Implementación

Elaboración de una “Carta de admisión”

- el *redactor* compone el documento en una lengua
- los *traductores* generan versiones en otros idiomas
- el *redactor* publica el documento multilingüe completo

Gestión de memorias de traducción

- complemento de otros gestores (WordFast, Déjà-Vu, etc...)

Trabajo cooperativo en red

- redactor y traductor comparten base documental
- recuperar documentos como punto de partida para la redacción y traducción de versiones actualizadas

Redactor multilingüe

- el *redactor* conoce una segunda lengua (con menor confianza)
- recupera un documento y modifica original y traducciones
- el *traductor* revisa y valida los cambios

Base de datos documental monolingüe

- el *redactor* recupera un documento y lo utiliza de plantilla

En general,

- agiliza el proceso de produc. de doc. multilingüe
- mejora la calidad y la cantidad de documentos generados

- ✓
 - Casos de uso
 - Descripción conceptual
 - Funcionamiento
 - Implementación

- Tratamiento de corpus
- Metadatos
- Usuarios y permisos

Primeramente contiene

- un corpus multilingüe, anotado, segmentado y alineado

Un documento

- se divide en subdocumentos (uno por lengua)
- etiquetados (estándar TEI)
- cada subdocumento está segmentado

existe alineamiento

- de los segmentos en las distintas lenguas

- ✓
 - Tratamiento de corpus
 - Metadatos
 - Usuarios y permisos

Categoría

- El más importante
 - clasifica el documento según una taxonomía jerárquica
 - dependiente del dominio
- Tres niveles
 - función (informar, inquirir y reglamentar)
 - género (25 géneros)
 - tema (256 temas)
- por ej. “certificado de asistencia a un cursillo”
 - informativo - certificado - asistencia a un cursillo

Estado

- Informa de la situación actual del documento
 - en lo referente a su multilingüismo
- Tres valores:
 - sin validar (texto inicial producido por el redactor)
 - validado (aprobado por los traductores)
 - normativo (versión multilingüe que sirve de modelo)

Visibilidad

- Grado de confidencialidad del documento
- Cuatro posibles valores:
 - borrador (visible solo para el redactor)
 - confidencial (visible con fuertes restricciones)
 - compartido (visible solo en la organización)
 - público (visible universalmente)

Centro

- Centro o departamento que origina el documento
- Separado en dos niveles:
 - centro
 - subcentro

Varias fechas:

- fecha original del documento
- fecha de inclusión en el corpus
- última modificación

- ✓ ● Tratamiento de corpus
- ✓ ● Metadatos
- Usuarios y permisos

Tipos de usuarios:

- **invitados**
 - externos a la organización
 - solo lectura
- **redactores**
 - añaden nuevos documentos
 - en una lengua o multilingües sin revisar
- **traductores**
 - realizan o revisan las traducciones
- **administradores**
 - gestionan el sistema

Permisos:

- . Usuarios
- . Metadatos:
 - . Visibilidad
 - . Propietario
 - . Estado

- ✓ ● Tratamiento de corpus
- ✓ ● Metadatos
- ✓ ● Usuarios y permisos

- ✓ ● Casos de uso
- ✓ ● Descripción conceptual
- Funcionamiento
- Implementación

- Inserción
- Modificación
- Visualización
- Recuperación por filtrado
- Recuperación por búsqueda textual
- Usuarios y permisos
- Ciclo de vida

El usuario aporta

- los metadatos
 - categoría y centro
 - fecha original
 - visibilidad y estado (sin validar)
- el texto de los subdocumentos (en cada lengua)

El sistema

- realiza el etiquetado,
- la segmentación y alineado
- y añade el documento al corpus



- Inserción
- Modificación
- Visualización
- Recuperación por filtrado
- Recuperación por búsqueda textual
- Usuarios y permisos
- Ciclo de vida

El usuario puede cambiar

- los metadatos
 - categoría y centro
 - fecha original
 - visibilidad y estado (según el tipo de usuario)
- el texto de los subdocumentos (en cada lengua)

El sistema

- realiza el etiquetado,
- la segmentación y alineado
- y actualiza el documento al corpus

- ✓ ● Inserción
- ✓ ● Modificación
- Visualización
- Recuperación por filtrado
- Recuperación por búsqueda textual
- Usuarios y permisos
- Ciclo de vida

Cuando se recupera un documento

- los contenidos segmentados y alineados de todas las lenguas,
- los contenidos sin segmentar
 - para facilitar la legibilidad
 - permitir copiar y pegar el texto completo
- se da acceso a la función de modificación de los metadatos

Se permite

- exportar cada subdocumento (por cada lengua) al formato TEI
- generar memorias de traducción (por cada par de lenguas) en formato TMX

- ✓ ● Inserción
- ✓ ● Modificación
- ✓ ● Visualización
- Recuperación por filtrado
- Recuperación por búsqueda textual
- Usuarios y permisos
- Ciclo de vida

Se ofrece un formulario

- metadatos estado, visibilidad, categoría, centro y corpus
- criterios de presentación
 - ordenar por centro, categoría, fecha o corpus
 - en orden normal (ascendente) o inverso

Por medio de filtros

estado:

visibilidad:

categoría:

centro:

corpus:

ordenar por:

inverso:

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 | Inauquración exposición | 7 | es eu | informar / tarjeta de invitación / acto cultural | ConsejoGob | TMXtore | 2003/06/13 | 2001/07/28 | Editar |
| <input type="button" value="actualizar"/> | | | | | | | | | | |

- ✓ ● Inserción
- ✓ ● Modificación
- ✓ ● Visualización
- ✓ ● Recuperación por filtrado
- Recuperación por búsqueda textual
- Usuarios y permisos
- Ciclo de vida

Búsqueda en texto libre

Texto de búsqueda:

en los idiomas: ▼

Buscar

Resultados de la búsqueda en segmentos

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

1 - [Reqlamento sobre fotocopias](#)

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

2 - [Reqlamento sobre fotocopias](#)

| | |
|-----------|------------------------------------|
| es 006 | Título del Libro o Revista: |
|-----------|------------------------------------|

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

| | |
|-----------|--|
| es 001 | El Instituto de Derechos Humanos Pedro Arrupe de la Universidad de Deusto le invita a la presentación del libro "El caso Awas 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 Awas Tigni, y de Mikel Berraondo, investigador del Instituto de Derechos Humanos Pedro Arrupe. |
| eu 001 | Deustuko Unibertsitateko Pedro Arrupe Giza Eskubideen Institutuak "El caso Awas Tingni contra Nicaragua: nuevos horizontes para los derechos humanos de los pueblos indígenas" liburuauren aurkezpena gonbidatzen zaitu. Ekitaldia maiatzaren 6an, asteartean, izango da arratsaldeko 7etan Deustuko Unibertsitateko Hitzaldi Aretoan eta James Anaya, Arizonako Unibertsitateko Nazioarteko Zuzenbideko katedraduna eta Awas Tigni komunitatearen lege aholkularia, eta Mikel Berraondo, Pedro Arrupe Giza Eskubideen Institutuko ikertzailea izango dira bertan. |

- ✓ ● Inserción
- ✓ ● Modificación
- ✓ ● Visualización
- ✓ ● Recuperación por filtrado
- ✓ ● Recuperación por búsqueda textual
- Usuarios y permisos
- Ciclo de vida

Los invitados

- Solo pueden ver documentos
 - con *visibilidad* “pública”
 - y *estado* “validado” o “normativo”

Los redactores

- pueden ver todos los documentos
 - excepto aquellos con *visibilidad* “borrador” o “confidencial”
 - de los que no sean propietarios
- pueden añadir y modificar documentos
 - excepto el metadato *estado*

Los traductores

- pueden ver todos los documentos
 - excepto aquellos con *visibilidad* “borrador”
 - de los que no sean propietarios
- pueden modificar documentos
 - excepto el metadato *visibilidad* en los documentos de los que no sean propietarios
 - se encargan de cambiar el metadato *estado*
- también pueden añadir documentos como cualquier redactor

Los administradores

- no tienen ninguna restricción

- ✓ ● Inserción
- ✓ ● Modificación
- ✓ ● Visualización
- ✓ ● Recuperación por filtrado
- ✓ ● Recuperación por búsqueda textual
- ✓ ● Usuarios y permisos
- Ciclo de vida

Un ciclo podría ser el siguiente:

- un redactor añade un nuevo documento
 - el redactor es su propietario
 - su *estado* será “sin validar”
 - y su *visibilidad* “borrador”
- cuando el redactor completa el contenido (una o más lenguas) del documento
 - su *visibilidad* será “confidencial”, “compartido” o “público”
- un traductor accede al documento
 - edita el texto, traduciendo el contenido
 - su *estado* será “validado”
 - y avisa al redactor original de que el documento ya es multilingüe

- ✓ ● Inserción
- ✓ ● Modificación
- ✓ ● Visualización
- ✓ ● Recuperación por filtrado
- ✓ ● Recuperación por búsqueda textual
- ✓ ● Usuarios y permisos
- ✓ ● Ciclo de vida

- ✓ ● Casos de uso
- ✓ ● Descripción conceptual
- ✓ ● Funcionamiento
- Implementación

Zope

- es una plataforma libre para desarrollar aplicaciones web
 - manejo óptimo de la información
 - gestión de usuarios
 - módulo Localizer

- ✓ ● Casos de uso
- ✓ ● Descripción conceptual
- ✓ ● Funcionamiento
- ✓ ● Implementación

/

- ✓ ● Introducción
- ✓ ● SARE-Bi
- Lineas futuras

X-Flow es el nuevo proyecto del grupo DELi

- sus resultados pueden ayudar a automatizar en gran medida la gestión del flujo de documentos multilingües

El objetivo general de X-Flow es facilitar

- la gestión del *flujo de contenidos multilingües*
- en el desarrollo de proyectos de traducción y localización
- en sistemas de información y publicación en Internet

En concreto, aportar una solución adecuada que complemente

- a los productos como `Localizer`
- en la gestión del flujo de documentos multilingües
- basándose en los estándares de intercambio
 - TMX y
 - XLIFF

En los últimos años han ido apareciendo nuevos estándares en el mundo de la localización

- XLIFF, XML-based Localization Interchange File Format
 - nace para recoger el testigo de TMX
- incorporando nuevas facilidades para la gestión de flujo o el control de versiones

- Cada documento estará formado por un conjunto de *versiones*
- y cada versión por varios *textos* en distintos idiomas
- Cada texto estará en un *estado de publicación*:
 - pendiente de corrección
 - pendiente de moderación
 - rechazado
 - oculto
 - o visible
- y en un *estado de traducción*:
 - pendiente de traducción
 - pendiente de revisión
 - o visible

/

- ✓ ● Introducción
- ✓ ● SARE-Bi
- ✓ ● Lineas futuras