RESUMEN

Éste Proyecto tiene como finalidad diseñar e implementar una herramienta que permita el acceso a la documentación que se genera en los **Congresos de Ingeniería de la Organización** (**CIOs**) organizados por **ADINGOR** de manera cómoda y sencilla; a la vez que segura. Esta documentación está formada por artículos de investigación en formato **PDF**.

La herramienta a desarrollar debe tener un acceso restringido. Donde una vez correctamente autenticado, el usuario podrá acceder a las opciones que permitan el procesado de nuevos documentos.

Para añadir un nuevo artículo; el primer paso será seleccionar el correspondiente archivo en formato **PDF** y la regla de parseo adecuada. A continuación, y ya de forma automática y transparente para el usuario, la herramienta extraerá el texto del documento en **PDF**. Sobre el texto extraído actuarán las reglas de parseo para obtener los datos necesarios para clasificar correctamente el artículo. Finalmente la información obtenida debe presentarse al usuario para que este la verifique y la corrija si fuera necesario.

Durante las etapas iniciales del Proyecto se platearon diferentes aproximaciones que pudieran dar respuesta a los requisitos definidos. Se valoró la posibilidad de desarrollar una aplicación web completamente nueva, partiendo de cero o dar continuidad a una ya existente. Incluso se realizó el análisis de alguna herramienta de gestión documental. Finalmente se decidió trabajar para mejorar una aplicación web ya existente.

Las tecnologías usadas para la realización de este Proyecto quedaron definidas cuando se decidió dar continuidad a la aplicación web original. La aplicación original se desarrolló usando **SYMFONY** como framework **PHP**, **Doctrine** como herramienta **ORM** y **MySQL** como base de datos.

Para extraer en forma de texto plano la información contenida en un archivo en formato **PDF**, inicialmente se pensó en usar el componente **PDFtotext** del proyecto de código abierto **XPDF**. Sin embargo, a pesar de los buenos resultados obtenidos, tuvo que ser desestimado porque la forma de invocar dicha herramienta suponía un grave riesgo para la seguridad de la aplicación y del servidor. Finalmente se optó por una librería implementada en **PHP**, la librería **PDFParser**.

Se ha estimado en **50** el número de jornadas necesarias para las diferentes fases del Proyecto. Con una tarifa de 40€ por hora para el jefe de proyectos (con una dedicación del 30%) y de 32,5€ por hora para el programador (con una dedicación del 100%), el coste de contratar a una empresa especializada para realización de este Proyecto se estima en **21.538€**.

El trabajo realizado en este Proyecto se puede dividir en tres grandes bloques. El primero se dedicó a la adaptación de la aplicación web original para reflejar los cambios de la base de datos. La base de datos sobre la que se construyó la aplicación web original había sido modificada. Se habían incluido nuevas tablas y nuevos campos donde almacenar información, que originalmente no se había tenido en cuenta. Se modificó el modelo de datos y se hicieron los cambios necesarios para que la información manejada por la aplicación quedara equiparada con la de la base de datos. Además, se actualizó el framework **SYMFONY** de la versión 1.3 a la 1.4.

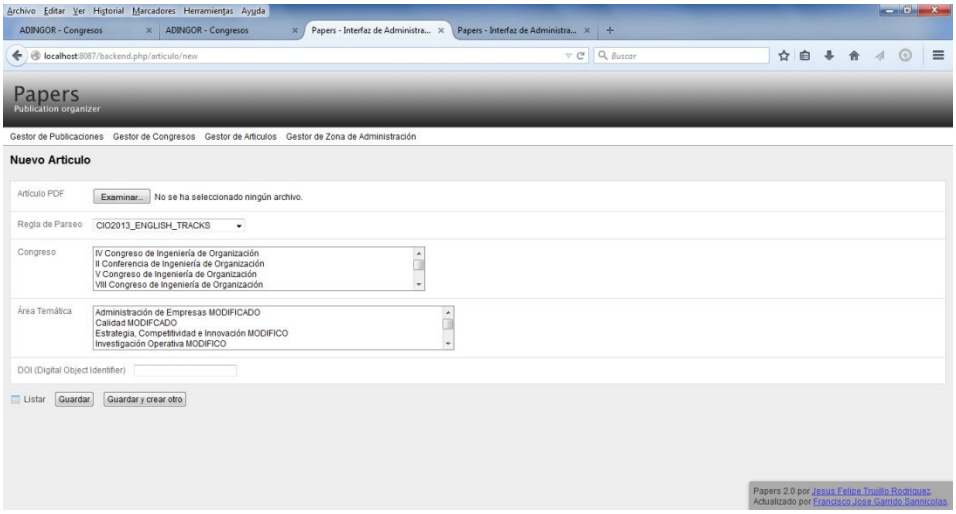

Figura 1. Formulario para añadir un nuevo artículo.

A continuación, se adaptó la aplicación web original para implementar el procesado automático de nuevos artículos.

Hasta ahora para añadir un nuevo artículo se debía seguir un proceso de dos pasos. El primero se realizaba de forma independiente a la aplicación y consistía en procesar el documento **PDF** para obtener un archivo **JSON** con los datos necesarios para clasificar correctamente el artículo. A continuación, de vuelta a la aplicación web, se daba de alta un nuevo artículo incorporando ambos archivos, tanto el **PDF** como el **JSON**. Gracias a las modificaciones realizadas estos dos pasos se reducen a uno solo. De esta forma se puede añadir un nuevo artículo directamente usando la aplicación web. Se definió el formulario apropiado para añadir el documento en **PDF** y seleccionar la regla de parseo adecuada y todo el procesamiento tanto previo como posterior al parseo.

El tercer bloque se dedicó al parseo automático. El objetivo del parseo automático es localizar en el texto extraído del artículo los datos que lo caracterizan.

El sistema permite buscar las siguientes características de un artículo: el título, el resumen, el título en el idioma secundario, el resumen en el idioma secundario, los autores y sus direcciones de correo electrónico, las palabras clave tanto en el idioma principal del artículo como en el idioma secundario, el número de la primera página del documento, el idioma principal y por último, el idioma secundario.

Todos estos datos deben ser localizados en el texto, excepto el idioma principal y el idioma secundario que son definidos por la propia regla de parseo.

Para seleccionar el texto deseado se usan funciones como **preg_match_all** y expresiones regulares. Esta función busca dentro del texto todas las coincidencias con la expresión regular indicada.

Para ayudarnos a evaluar la aplicación se analizaron los resultados obtenidos durante la fase de pruebas. En la siguiente tabla se puede ver un resumen de los resultados obtenidos con las reglas de parseo implementadas para los Congresos de los años 2013 y 2014; para artículos en español y en inglés, tanto "abstracts" (resúmenes) como comunicaciones completas.

Regla de parseo	Datos localizados	Errores	Porcentaje
CIO2013_ENGLISH_TRACKS	30	2	7%
CIO2013_SPANISH_TRACKS	45	16	36%
CIO2014_FULL_PAPER_ENGLISH	30	0	0%
CIO2014_FULL_PAPER_ESPAÑOL	45	0	0%
CIO2014_EXTENDED_ABSTRACTS	25	2	8%
TOTAL	**175**	**20**	**11,5%**

Tabla 1. Resumen de los resultados obtenidos.

De los 175 datos obtenidos automáticamente durante el parseo automático, 20 se encontraron erróneos. Estimando que casi el 90% de los datos se identifican correctamente, con el consiguiente beneficio que esto implica.

Además, conviene resaltar que si no se tienen en cuenta los resultados obtenidos con la regla de parseo CIO2013_SPANISH_TRACKS el porcentaje de datos erróneos sería menor al 4%. Esta regla de parseo genera muchos errores como consecuencia de un error en el formato de la plantilla usada.

Estos datos nos llevan a dos conclusiones: la primera es la confirmación del buen funcionamiento de la aplicación y, en concreto del parseo automático; y la segunda es la importancia de que los autores de los artículos tomen conciencia de las ventajas que aporta el correcto uso de las plantillas que se proporcionan en los congresos.