Un sistema de administración del flujo de trabajo de creación de repositorios documentales para bibliotecas digitales[1]

José A. Cotelo	José R. Paramá	Miguel R. Penabad	Ángeles S. Places	Eloy V. Fontenla
Dep. de Computación	Dep. de Computación	Dep. de Computación	Dep. de Computación	Dep. de Computación
Fac. de Informática	Fac. de Informática	Fac. de Informática	Fac. de Informática	Fac. de Informática
Univ. da Coruña	Univ.da Coruña	Univ.da Coruña	Univ.da Coruña	Univ.da Coruña
Campus de Elviña	Campus de Elviña	Campus de Elviña	Campus de Elviña	Campus de Elviña
15071 A Coruña	15071 A Coruña	15071 A Coruña	15071 A Coruña	15071 A Coruña
jaclema@mail2.udc.es	parama@udc.es	penabad@udc.es	asplaces@udc.es	evazquez@udc.es

Resumen

La creación de un repositorio documental para bibliotecas digitales es un proceso complejo y, como tal, propenso a errores. Cada documento, normalmente en formato impreso, debe ser procesado en una serie de pasos que incluyen su digitalización, el reconocimiento óptico de caracteres, corrección de los posibles errores, indexación y posterior publicación en web. Además, es habitual que en este trabajo intervengan varios equipos de personas realizando tareas de diferente naturaleza, a las que hay que coordinar para asegurar que el proceso global se realiza de la forma más eficiente posible. Por ello, es imprescindible contar con un sistema que gestione y coordine todos los recursos que intervienen en el proceso, guiando la realización de cada actividad y automatizando en lo posible tareas (como el nombrado de los ficheros en los que se van almacenando los resultados intermedios del proceso) muy tediosas y foco de errores. En este artículo se presenta DigiFlow, un sistema de administración del flujo de trabajo implícito en la creación repositorios documentales, y su implementación para la construcción del repositorio documental de la Hemeroteca Virtual de la Real Academia Galega (RAG).

1. Introducción

En los últimos años se ha dedicado mucho esfuerzo a la publicación de documentos de todo tipo en Internet, normalmente en bibliotecas digitales. Un tipo especial de documentos son aquellos que sólo existen en versión papel y que no son fáciles de encontrar ya que suelen estar almacenados en bibliotecas de difícil acceso, incluso descatalogados, y muchas veces presentan un mal estado de conservación. Este es el caso de documentos antiguos de los siglos XVI-XVIII, tales como los Libros de Emblemas [12] o las Relaciones de Sucesos [13]. Documentos de este tipo, probablemente, no serán nunca reeditados en papel debido a que no resulta viable desde el punto de vista económico. La construcción de una biblioteca digital para estos documentos tiene dos propósitos: contribuir a su preservación y hacerlos disponibles para un público más amplio.

De entre los retos implicados en la construcción de bibliotecas digitales:

- la especificación y prototipado rápido de bibliotecas digitales [3][4],
- compresión de textos [5][6],
- el desarrollo de servicios para bibliotecas digitales [1] o
- la creación de repositorios documentales para bibliotecas digitales [2],

en este artículo nos centramos en el último de ellos, es decir, en el proceso de alimentar una biblioteca digital, tanto con metadatos sobre los documentos, como con los documentos en sí mismos (tanto en formato imagen como en texto plano para permitir búsquedas por contenido). En este proceso es necesario llevar a cabo una serie de tareas en una secuencia determinada. Una posible secuencia se muestra en la Figura 1. Primero, se almacenan en la biblioteca digital los

Tabla 1.

[1] Este trabajo está parcialmente financiado por el MCYT (PGE y FEDER) ref. TIC2003-06593.

metadatos sobre los documentos (probablemente en una base de datos documental o quizás en ficheros XML). Después de este paso, los documentos pueden ser escaneados página a página y se puede obtener el texto de cada una de ellas aplicándoles un OCR (Reconocedor óptico de caracteres). Aunque los OCR actuales han mejorado mucho con respecto a los primeros que se comercializaron, todavía siguen teniendo problemas al reconocer el texto de documentos antiguos ya que en ellos las tipografías no suelen ser las estándar o resultan difíciles de reconocer debido al estado de conservación de los documentos. Por esta razón es necesario revisar los textos que automáticamente extrae el OCR de las páginas digitalizadas de los documentos. Finalmente, debe existir una fase en la que estos documentos se indexen y se publiquen, probablemente, en Web.

Para facilitar el proceso de creación de un repositorio documental para una biblioteca digital, hemos definido formalmente el flujo de trabajo descrito en el párrafo anterior, y hemos desarrollado DigiFlow, un sistema de administración del flujo de trabajo de creación de repositorios documentales que permite realizar todas estas tareas de la manera más eficiente posible, mejorando tanto la calidad del texto y las páginas que se obtienen como resultado después de todo el proceso, como los costes en tiempo y recursos del proceso global.

En este artículo describimos las funcionalidades de DigiFlow, así como su primera aplicación para la construcción del repositorio documental para la Hemeroteca Virtual de la Real Academia Galega accesible en http://www.realacademiagalega.org/Hemeroteca.

Una vista general de la arquitectura de DigiFlow se muestra ya en la Figura 1. En su diseño hemos seguido, tanto las buenas prácticas de desarrollo de proyectos de sofware, como las recomendaciones dadas en el *WFMC Reference Model* [7] publicado ya por la WFM Coalition [8][9], la cual sigue trabajando en la definición de estándares para sistemas de workflow.

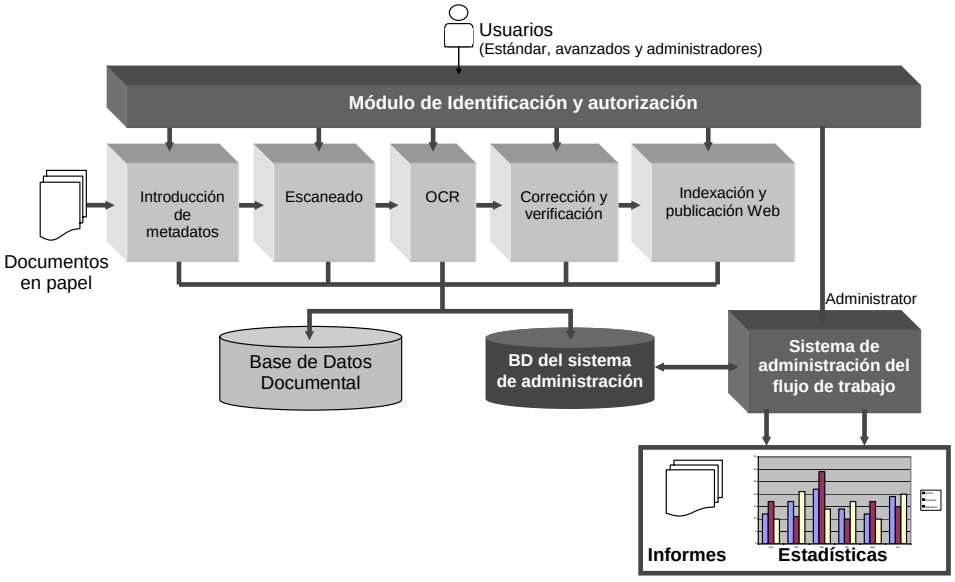

Figura 1. DigiFlow, un sistema de administración del flujo de trabajo de creación de repositorios documentales

El resto del artículo se organiza como sigue: la sección 2 describe las actividades que forman el flujo de trabajo que permite administrar DigiFlow y las funcionalidades de esta herramienta; la sección 3 describe algunos aspectos de su uso en la creación de la Hemeroteca Virtual de la Real Academia Galega; y, finalmente, la sección 4 introduce nuestras conclusiones y líneas de trabajo futuras.

2. DigiFlow

2.1. Definición del flujo de trabajo

Como ya se ha comentado, DigiFlow es un sistema diseñado para gestionar el flujo de trabajo que hay que realizar para la obtención de una réplica digital de documentos en papel. En este artículo llamamos *obras* a los documentos. Una obra puede ser un libro, una revista o cualquier otra unidad que pueda ser digitalizada.

El flujo de actividades necesarias para procesar cada obra se muestra en la Figura 2. Una actividad se puede ver como una fase del flujo de trabajo.

Las actividades pueden ser llevadas a cabo a través de diferentes aplicaciones, pero deben tener una interfaz común que facilite el trabajo de los usuarios encargados de realizarlas. Además las actividades pueden ser divisibles o indivisibles. Una actividad divisible puede ser realizada por varios usuarios en paralelo (por ejemplo, dos usuarios pueden escanear a la vez distintas páginas de una misma obra). En este caso, la actividad se divide en tareas. Una tarea es una implementación de una actividad que debe ser realizada únicamente por un usuario. Como se muestra en la Figura 2, para cada obra deben realizarse las siguientes actividades:

- *Comienzo del flujo de trabajo con una obra:* Esta actividad es la primera del flujo de trabajo. Se trata de "dar de alta" la obra en el sistema y de asignar la tarea de introducción de metadatos a un usuario.
- *Almacenamiento de metadatos*: Esta actividad consiste en alimentar el sistema con los metadatos relativos a una obra. Llamamos metadatos a información como el título, el autor, la fecha y el lugar de publicación, etc. Como se muestra en el diagrama de la Figura 2, no es posible continuar con el flujo de trabajo hasta que los metadatos no estén correctamente almacenados. Esto es así, porque los metadatos son información indispensable para crear y distribuir el resto de tareas del proceso entre las personas encargadas de realizarlas.
- *Asignación de tareas:* Esta actividad consiste en la generación de las actividades y tareas necesarias para completar el flujo de trabajo. En el caso de una actividad divisible, el sistema permite generar tareas diferentes y asignarlas a diferentes usuarios del sistema de manera que puedan ser llevadas a cabo en paralelo.

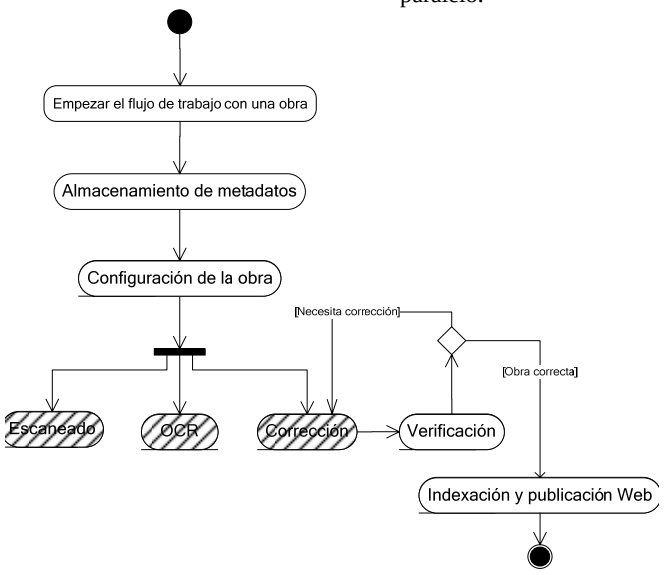

Figura 2. Definición del flujo de trabajo de creación de un repositorio documental.

- *Escaneado:* Esta actividad consiste en la creación de las imágenes digitales correspondientes a cada página de una obra. El sistema genera el directorio y los nombres de los archivos en los que guardar las imágenes de forma automática, sin que, en ningún caso, sea necesario la intervención del usuario.
- *OCR:* Usando una aplicación de OCR que puede ser comercial o desarrollada ad hoc, esta actividad procesa las imágenes generadas durante la actividad anterior para cada una de las páginas de una obra. Esta actividad se lleva a cabo de forma automática, pero a veces es necesario, manualmente, delimitar áreas especiales como zonas demasiado deterioradas en las que es prácticamente imposible reconocer los caracteres.
- *Corrección:* Las herramientras de OCR no siempre proporcionan los resultados esperados, especialmente si la obra no tiene una tipografía estándar, si la calidad del papel no es la óptima o si el estado de conservación de la obra no es bueno. Por eso, a pesar de que el OCR se puede entrenar para hacer algunas correcciones de forma automática, usando diccionarios, no es posible dejar de hacer una cuidadosa comprobación del resultado para verificar su corrección y enmendar los posibles errores.
- *Verificación:* Aunque el usuario es guiado por el sistema en el sentido de que nunca tiene que recordar cuál es la siguiente página a escanear, por ejemplo, y, para cada tarea que realiza, puede ver el resultado antes de confirmar su correcta realización al sistema, es posible que, por error, se hayan escaneado páginas que no se correspondían con las que pedía el sistema o que se haya escaneado una página en diferente sentido al que ordena el sistema. Por eso, esta actividad es una segunda revisión de algunas de las páginas para verificar la corrección de todo el proceso (escaneado, OCR y corrección).
- *Indexación y publicación web de las páginas:* Estas actividades incluyen la actualización de los índices utilizados para permitir posteriormente una búsqueda de documentos por contenido y la publicación Web de las obras totalmente procesadas.

El diagrama de la Figura 2 indica que el escaneado, OCR y corrección son actividades paralelas a nivel de obra, es decir, se pueden realizar simultáneamente para una obra concreta, pero, por supuesto, una página concreta de una obra no puede ser escaneada y corregida al mismo tiempo.

2.2. Funcionalidades del sistema

Las funcionalidades que se encuentren disponibles en DigiFlow dependen del perfil del usuario que se conecte al sistema. Se distinguen tres perfiles de usuario:

- *Usuario estándar:* puede realizar cualquiera de las actividades de escaneado, OCR y corrección descritas, pero no tiene responsabilidad alguna en la gestión del sistema.
- *Usuario avanzado:* es responsable de actividades críticas como almacenamiento de metadatos o verificación de la corrección de una tarea. Un usuario avanzado es habitualmente un especialista en filología y humanidades.
- *Administrador:* este tipo de usuario puede realizar las tareas específicas de los otros dos perfiles y además crear nuevos usuarios, cambiar el pérfil de un usuario, empezar el flujo de trabajo para una nueva obra, asignar tareas y monitorizar el proceso global.

La funcionalidad básica del sistema permite que la realización de las actividades del flujo de trabajo sea mucho más simple ya que proporciona un entorno integrado para realizar cada una de ellas. Una vez que el usuario se identifica, accede a una tabla en la que puede ver su lista de tareas pendientes ordenada por prioridades (ver Figura 3). La ejecución de cada una de estas tareas se realiza a través de una aplicación específica dependiendo del tipo de actividad de que se trate.

En la Figura 3 se muestra una pantalla en la que se está realizando el escaneado de una obra. El sistema le está diciendo al usuario exactamente qué página tiene que poner en el escáner antes de pulsar el botón *escanear*. El usuario no tiene que preocuparse de llevar la cuenta de qué página le toca escanear, o con qué nombre tiene que guardar el archivo cuando termine de escanearla. Este tipo de cuestiones, tan propensas a errores, son tratadas

de forma autormática por el sistema de administración del flujo de trabajo.

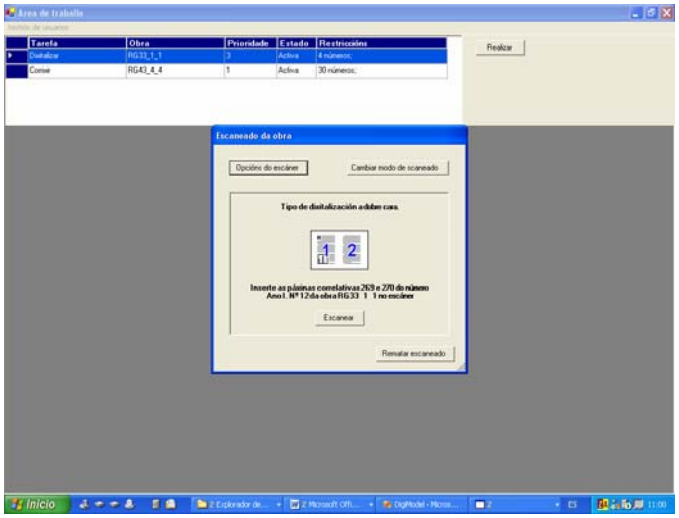

Figura 3. Área de trabajo de un usuario

DigiFlow también ofrece algunas funcionalidades dirigidas a monitorizar y controlar el flujo de trabajo de manera que sea posible actuar para mejorar el resultado del proceso global. Son los usuarios administradores los que tienen acceso a estas funcionalidades. Algunos de los parámetros que se pueden conocer a través de DigiFlow son:

- El reparto de tareas actual entre los usuarios estándar o avanzados del sistema.
- El estado actual de procesamiento de cada obra.
- El número de horas que se tardó en realizar una actividad para una obra determinada.

Los administradores pueden cambiar la prioridad de una tarea o pueden modificar la asignación inicial de tareas a usuarios. De esta manera pueden priorizar el procesamiento de determinadas obras, pueden evitar cuellos de botella en alguna actividad, o solucionarlos si ya se han producido.

En la Figura 4, aparece la cabecera de uno de los informes que se pueden obtener con DigiFlow. En este caso se trata de un informe que permite analizar la carga de trabajo de los usuarios. El gráfico usa un código de colores para indicar cuantas tareas de cada tipo tienen asignadas los usuarios. Debajo de este gráfico se describen detalladamente las tareas de que se trata.

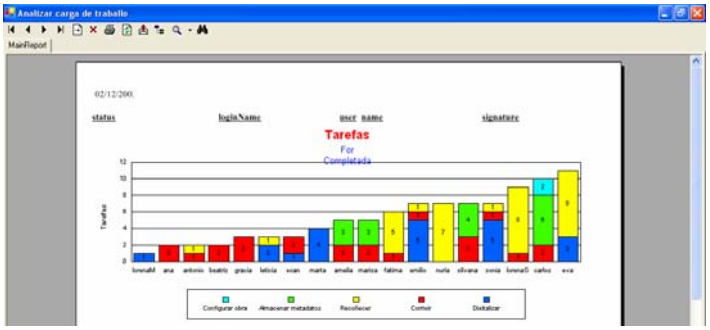

Figura 4. Informe con el tipo y número de tareas asignadas

2.3. Tecnología de desarrollo

El diseño de la arquitectura de DigiFlow sigue las recomendaciones de la *Workflow Management Coalition (WFMC)*, así como una metodología formal de desarrollo, el Proceso Unificado de Desarrollo de Software. El uso de patrones de diseño, así como *Value Object*, *Data Access Object*, *Abstract Factory*, *Session Facade*, *Business Delegate*, *Model View Controller* y *Layers*, también fueron de gran ayuda para conseguir un diseño robusto y escalable de la aplicación.

En cuanto a su implementación, DigiFlow fue desarrollada usando C#, principalmente, de la plataforma MS Visual Studio .NET. Como Sistema Gestor de Bases de Datos (SGBD) para la base de datos de administración hemos elegido Microsoft SQL Server 7.0, porque en nuestro grupo de investigación [14] ya se está usando este gestor para otros proyectos y sabemos que es lo suficientemente robusto y eficiente como para dar soporte a esta herramienta. De todos modos, DigiFlow ha sido construido para soportar cualquier otro SGBD tan sólo realizando pequeños cambios en la configuración del sistema y sin necesidad de reprogramar ningún módulo.

3. Aplicación de DigiFlow para la creación de la Hemeroteca Virtual de la RAG

DigiFlow ha sido diseñado para atender a un flujo de trabajo genérico como el que se muestra en la Figura 1. Sin embargo, la motivación inicial del sistema fue su aplicación a la digitalización de los fondos de la Real Academia Galega (RAG) [11], dentro del marco de un convenio de colaboración con la Universidade da Coruña (en concreto, con el Laboratorio de Bases de Datos [14]) para crear su Hemeroteca Virtual.

La Real Academia Galega cuenta con una Hemeroteca de revistas del siglo XIX, la mayor parte escritas en español y algunas de ellas en gallego. Estas obras tienen un valor incalculable ya que permiten estudiar la situación de Galicia en los últimos siglos. Debido a su antigüedad y al estado de conservación de los ejemplares (ver Figura 5) estas revistas no están disponibles para consulta para el público en general. Por eso, para facilitar el acceso a las mismas, la RAG decidió crear una Hemeroteca Virtual accesible a través de Internet.

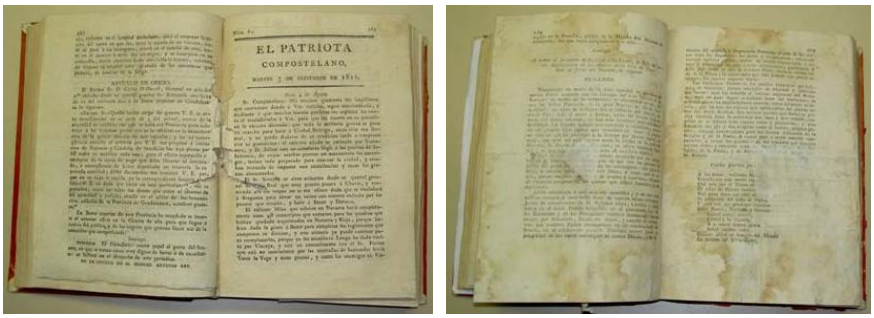

Figura 5. Imagen de *El Patriota compostelano* (1810)

En esta aplicación de DigiFlow, una obra es cada uno de los volúmenes encuadernados que contienen números de revistas. Cada volumen contiene varios números de revistas, los cuales pueden pertenecer a revistas totalmente distintas. Como se muestra en la Figura 6 (Diagrama Entidad-Relación de la base de datos documental de la Hemeroteca), cada número contiene varios artículos de varias páginas, cada una de las cuales debe ser escaneada, transformada en texto y corregida. Puede ocurrir que en una misma página finalice un artículo y comiencer el siguiente.

El flujo de trabajo para la creación del repositorio documental de esta hemeroteca tiene dos variantes, diferentes entre sí en la primera actividad del flujo, el almacenamiento de metadatos. En un caso, los metadatos de las revistas son introducidos manualmente en el

sistema, mientras que para algunas obras pueden ser importados de forma automática del catálogo de revistas de la RAG. La base de datos que almacena los metadatos ha sido creada usando Microsoft SQL Server 7.0 y las aplicaciones de introducción de metadatos fueron desarrolladas en C# de la plataforma MS Visual Studio .NET, con una interfaz bien definida que permitió su integración en DigiFlow.

La aplicación que se eligió para realizar el escaneado y OCR de las obras fue Scansoft OmniPage Pro 12 [10] que cuenta con una interfaz de automatización OLE *(OLE Automation Interface*), sencilla de utilizar.

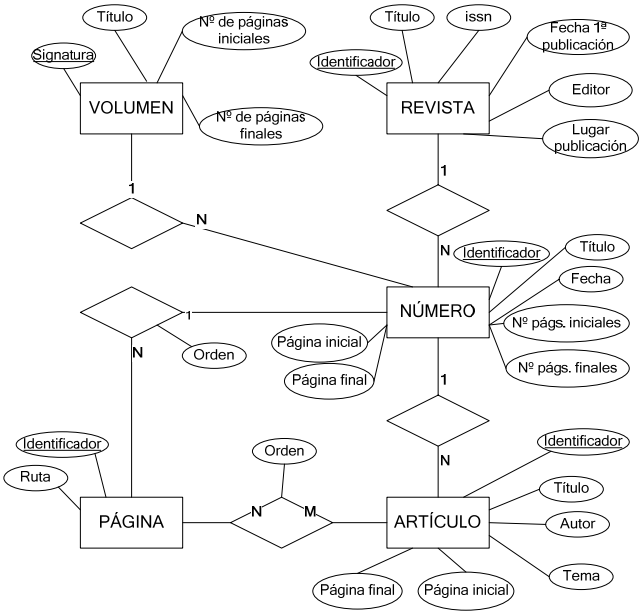

Figura 6. Diagrama Entidad-Relación de la base de datos documental

En este proyecto participaron simultáneamente en alguna de las actividades del flujo de trabajo una media de 20 personas, estudiantes, o recién licenciadas, de la Licenciatura en Filología Gallego-Portuguesa de la Universidade da Coruña. Estas personas estuvieron trabajando durante varias etapas entre octubre de 2003 y enero de 2005.

La Hemeroteca Virtual se creó con números de 27 revistas distintas sumando un total de 23.000 páginas. En la Tabla 1 se hace un resumen del tiempo consumido por algunas de las etapas del flujo de trabajo para la creación de este repositorio documental. Probablemente, el tiempo que se tardó en el escaneado de los documentos sea menor que el que se hubiese necesitado si los usuarios tuviesen la responsabilidad de nombrar cada uno de los ficheros, almacenarlos en el lugar adecuado o recordar qué páginas han escaneado ya y cuál es la siguiente que tiene que ser tratada. La actividad de corrección del OCR es la que más tiempo consume debido a la mala conservación del papel y a la tipografía de las revistas de la hemeroteca. Aunque el OCR se ha entrenado durante el proceso sigue introduciendo errores, porque la naturaleza de los mismos va en función de cada revista.

Tabla 1. Estado actual del trabajo

Actividades	Páginas	Horas	Págs/ Hora
Almacenamiento de los metadatos	23.000	235,09	97,83
Escaneado	23.000	442,17	52,01
OCR	23.000	668,99	34,38
Corrección	21.751	7.854,89	2,77

Actualmente se están corrigiendo las últimas páginas, se está verificando la corrección de cada obra y también se está depurando la aplicación web de la Hemeroteca Virtual. De hecho, la mayor parte de las revistas pueden consultarse en http://www.realacademiagalega.org/Hemeroteca. Esta Hemeroteca Virtual es pionera en las búsquedas por el contenido que implementa, ya que devuelve los documentos que se ajustan a la consulta marcando en las propias páginas digitalizadas (imágenes) las palabras buscadas.

4. Conclusiones y trabajo futuro

La creación de un repositorio documental no es un proceso simple. Aunque el número de actividades diferentes puede parecer pequeño, el gran número de documentos y de personas implicados en su procesamiento es grande y, por lo tanto, es susceptible de errores. El tener tener un sistema que coordine y controle todo el proceso y que automatice las tareas más tediosas es imprescindible.

En este trabajo hemos presentado DigiFlow, un sistema de administración del flujo de trabajo de creación de repositorios documentales. El primer prototipo implementado de DigiFlow ha sido usado para un caso real, en el marco de un convenio de colaboración que la Universidade da Coruña mantiene con la Real Academia galega, pero la herramienta ha sido diseñada para administrar el flujo de trabajo general que hemos descrito.

Como trabajo futuro, queremos implementar las interfaces de acceso a otras aplicaciones comerciales de escaneado, OCR y almacenamiento de metadatos.

Referencias

[1] Brisaboa, N. R., Durán, M. J., Penabad, M. R., Places, A. S. A Collaborative Framework for a Digital Library. VI International Workshop on Groupware (CRIWG'2000). IEEE Computer Society Press (ISBN: 0-7695-0828-6), pp. 104-111. Madeira Island, Portugal. 2000.

[2] Jesús González Lorca, José Vicente Rodríguez Muñoz. *La tecnología del flujo de trabajo en el contexto de la biblioteca digital*, Anales de documentación, nº 5, pp. 157-175, 2002. [http://www.um.es/fccd/anales/ad05/ad0508.pdf]

[3] Lourdes Fernández, J. Alfredo Sánchez, Alberto García: MiBiblio: personal spaces in a digital library universe. Proceedings of the *Fifth ACM Conference on Digital Libraries*, June 2-7, 2000, San Antonio, TX, USA. ACM 2000, pages: 232-233

[4] Rohit Kelapure, Marcos André Gonçalves, Edward A. Fox: Scenario-Based Generation of Digital Library Services. Research and Advanced Technology for Digital Libraries, 7th European Conference, ECDL 2003, LNCD 2769 pages: 263-275.

[5] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman, May 1999.

[6] N. Brisaboa, E. Iglesias, G. Navarro, and J. Paramá. An efficient compression code for text databases. *In 25th European Conference on IR Research, ECIR 2003*, LNCS 2633, pages 468–481, 2003.

[7] Hollingsworth. WFMC Reference Model. November 1994. [www.wfmc.org/standards/docs/tc003v11.pdf]

[8] Aalst, Wil van der. *Workflow management: models, methods, and systems*. Cambridge, Massachusetts. The MIT Press 2002

[9] WFMC Web page [http://www.wfmc.org]

[10] ScanSoft OmniPage Pro Web page. [http://www.scansoft.com/omnipage]

[11] Portal Web da Real Academia Galega. [http://www.realacademia.org]

[12] Biblioteca Virtual de Literatura Emblemática. [http://rosalia.dc.fi.udc.es/emblematica-proyecto/index.html]

[13] Biblioteca Digital de Relaciones de Sucesos [http://rosalia.dc.fi.udc.es/Relaciones]

[14] Laboratorio de Bases de Datos [http://rosalia.dc.fi.udc.es/lbd]