Managing the workflow of massive feeding of digital libraries *

Ángeles S. Places, Nieves R. Brisaboa, José R. Parama,
Oscar Pedreira, Diego Seco

Database Laboratory, Facultade de Informatica, University of A Coruña
Campus de Elviña s/n, 15071 A Coruña, Spain
{asplaces,brisaboa,parama,opedreira,dseco}@udc.es

Abstract. Feeding a digital library requires a significant effort due to the cost of a massive digitization and processing of documents, specially when these documents are not in a good state of conservation. When a big group of people work together in this complex process, support tools are needed to manage the process workflow, to facilitate the work in each activity and thus, to reduce the cost and ensure the quality of the obtained results. This paper proposes a set of strategies to face the management of the workflow of the digital library building process and a general system architecture supporting them. The paper also presents a tool called DigiFlow developed following this architecture, and the results obtained from a real experience applying it in the digitization of a collection of 23,000 documents of the 19^{th} century.

Key words: Digital library, digitization, workflow management

1 Introduction

The interest and research on digital libraries has experienced an important growth during the last years, in parallel with the advances in document digitization, information retrieval and web publishing technologies. Among all the applications of these type of systems, one of the most important is the development of digital libraries that store and provide access to collections of ancient documents. The digitization, storage and indexing of these type of documents guarantees the correct preservation [1] of old texts that will not be reedited and that are in a critical state of conservation. In addition, electronic editions provide public access to a collection of documents that would be impossible otherwise [2]. The special problems that appear in this type of systems have opened interesting challenges in the development of digital libraries.

* This work has been partially supported by "Xunta de Galicia" grants PGIDIT05SIN10502PR and 2006/4, and by "Ministerio de Educación y Ciencia" (PGE y FEDER) grant TIN2006-15071-C03-03, and (Programa FPU) AP-2006-03214 (for O. Pedreira)

Advances in Artificial Intelligence and Applications
Research in Computer Science 32, 2007, pp. 352–362

Received 09/06/07
Accepted 31/08/07
Final version 30/09/07

Perhaps the main problem in digital libraries storing ancient documents is the importance and complexity of the digitization and processing of documents. These documents are usually very old and their state of conservation is, in general, poor. The characteristics of the documents can make necessary the use of special hardware for scanning. In addition, this activity has to be done with care to avoid new damages to the document. The conversion of the obtained images to text through character recognition technologies is specially difficult, due to the deterioration of the documents. Thus, the results of this task must be reviewed in order to correct possible errors. Other activities as metadata definition, document markup and text indexing are usually needed. Sometimes this process is carried out step by step by a small group of experts, and their skills and knowledge guarantee the quality of the results. But when the digital library has to be built from thousands of documents, the creation of the document repository involves a large digitization process carried out by a big group of people who are not always so skilled due to financial restrictions [3]. In these situations, the use of support tools to guide the workflow of the process and to facilitate the work of the people is mandatory.

The document repository building requires the ordered execution of a set of activities on the documents, with several people participating in each of them. In such a complex process, the lack of control on the workflow can result in dead times, errors in the obtained results and loss of data. In general, an unsatisfactory coordination of the people increases the overall cost and decreases the quality of the results. From previous experiences, we identified typical errors like the digitization of the same document several times with different names, errors according file naming conventions, publishing of no reviewed documents, loss of files, missing documents, etc. Because building a digital library requires significant effort, the more automated tools that can be built and used, the better will be the use of human resources [4]. The control of the workflow inside this work team is a key factor in the success of the digitization process. This control can be achieved by the use of a workflow management tool specially designed for this process. That is, a system which allows to coordinate and control all the involved people, monitor and manage factors as the current state of each page, store intermediate results, control the average time to process each document, record all the people who have worked in each document, etc. Previous works have developed support tools for the creation of digital libraries, but they do not consider the problems found in the digitization and process of a large collection of documents [5, 6].

This paper proposes a set of strategies for the workflow management of the repository creation process, and a general system architecture supporting them. The proposed strategies improve the performance of the process, ensure that all the necessary tasks are correctly performed and facilitate the work of the people devoted to this activity. In addition, we have implemented a tool called DigiFlow following this architecture. In this paper we also present the real experience of building a digital library for the *Royal Galician Academy*, from a collection of

23,000 documents of the 19^{th} century. We also present quantitative data about the results obtained in each step of the document repository building.

The rest of the paper is organized as follows. Next section analyzes the main problems found in the massive digitization and processing of documents for large digital libraries. Section 3 presents the requirements and the architecture to solve these problems. Section 4 presents the general architecture for workflow management in the digitization process. Sections 5 and 6 describe the implemented tool and the results obtained in its application to a real project. Finally, Section 7 presents our conclusions and directions of future work.

2 Problems in massive digitization of documents

One of the main problems found when building digital libraries is the complexity and cost of the digitization and processing of documents. Each page of a document has to be scanned, processed using OCR technologies for text extraction, corrected and reviewed. Other activities as metadata definition or document markup are also necessary. Finally the text is indexed, stored and published. Taking into account that the collection of documents can have hundreds of thousands of pages, the complexity and cost of document digitization is obvious. From our previous experience in the development of digital libraries, we have identified several typical problems in this process:

- Problems with the file naming protocol. Due to the high number of files to be managed during the digitization, such a protocol is necessary. When few people participate in the digitization, the file naming conventions are usually followed, and small errors can be easily managed. However, when tens of people are working together, small errors are likely to appear, and their management can produce a significant waste of time.

- Loss of files. Without support tools, each participant is the responsible of the files obtained in each activity. If the management of hundreds or thousands of files is done manually, typical errors will happen frequently. Overwriting files, saving files with the wrong name or in the wrong folder are a common source of lost files. If the experience of the participants with computers is limited, these errors will be very common.

- Task specification. There are different ways to carry out a task. A bad specification of the task parameters is also a source of typical problems. For example, scanning with an incorrect orientation of the pages, scanning two pages together instead of one, reviewing an already reviewed document or writing again the document metadata when they were already available in the database. This problem can be worst when several people work with the same document.

- Lack of coordination. Coordination is difficult when a big group of people work in the project. Each person can be devoted to specific activities and

have his/her own timetable. For example, a given person can be the responsible of scanning a document in the morning and other can be the responsible of correcting it in the afternoon. An effective task management to avoid dead times and waste of resources is necessary.

- Effective resource control. Since the number of resources used for the digitization is limited, the lack of control of them can be a source of dead times in some activities of the chain. For example, some workers could have to wait for free scanners or computers, or even for the availability of the physical document. In addition, without this resource management, reports about the particular resource used in each activity are not available.

- Responsibilities management. The correct definition of the responsible of each task is also important, specially when the review of the extracted texts is difficult and requires deep knowledge of the type of literature being digitized.

Perhaps most of these problems seem trivial and easy to solve. But taking into account that they can be repeated thousands of times during the whole digitization process, their consequences can have a really great impact in factors such as the process time and the quality of the digitized documents.

3 Requirements for the workflow management system

As a solution for the problems presented in the previous section, we describe here the workflow strategies we suggest for the workflow management architecture we propose in this paper.

- Automatized results management. As we mentioned in previous section, errors according the file naming conventions and loss of files are common in a group of people working together in a digitization chain. A workflow management system for document digitization should automatically manage the files produced in each activity. Thus, when an worker starts a new task, the system must automatically bring him/her the inputs needed for such a task (which are outputs of previous tasks), without any kind of human intervention. This support avoids problems of lost of intermediate products.

- Task specification. The system should offer to the administrator tools to continuously monitor different aspects of the workflow. Particularly the state of each document digitization task, which was assigned to a worker is of special interest, but also the progress of the results and possible problems recorded.

- Task control and monitoring. The system should offer to the administrator the tools needed to continuously monitor all the information about the state of each document task assigned to each worker. For example, the worker assigned to the task, the progress of the results and possible problems recorded.

- Effective resource management. This requirement is related with the previous one. The system should continuously control the availability of the necessary resources for each activity, identifying and informing immediately about possible conflicts between tasks due to the used resources. For example, if several documents are being scanned at the same time and a rescanning is needed to correct OCR errors, the system should identify a time slot in which the hardware will be available.

- Work dedication reporting. It is important to provide the possibility of generating reports about the average time devoted to each task, the average number of pages in a period of time, the number of corrections of the results of the OCR, the average dedication of each worker in a given period of time, etc.

- Product quality control. As we mentioned previously, when dealing with ancient documents, the review of the results obtained from the digitization can be really important due to the high error rate in the results of the OCR. Thus, the system must facilitate this review process, providing the reviewer with both the image and the extracted text without the need of obtaining the original document, and ensuring that the document is not published until the review is successfully finished.

4 System architecture

According to [7], workflow is concerned with the automation of procedures where documents, information or tasks are passed between participants following a defined set of rules to achieve, or contribute to an overall business goal; the computerized facilitation or automation of a business process, in whole or part. Workflow management systems can be classified in several types depending on the nature and characteristics of the process [8, 9]. Collaborative workflow systems automatize business processes where a group of people participate to achieve a common goal. This type of business processes involves a chain of activities where the documents, which hold the information, are processed and transformed until that goal is achieved. As the problematic of document repository building fits perfectly in this model, we based the architecture of the system in this model.

In general, we can differentiate three user profiles involved in the repository building:

- Administrator. Administrators are the persons responsible of the digitization process as a whole. They are the responsible of assigning tasks to different workers and controlling the state of each digitized document.

- Advanced users. The advanced users are the persons in charge of carrying out critical activities such as the metadata storage or the review of the texts

extracted from the OCR processes. The rationale behind this user type is that they usually need deep knowledge of the documents for performing these tasks (for example, a deep knowledge of the *Galician* literature of the 19^{th} century is needed if the user is going to review this type of documents).

- Standard users. The standard users are the workers who carry out tasks as scan or the OCR correction. This role is played by users with some knowledge in the document field but without any responsibility on the management of the system (for example, granted students could carry out these activities).

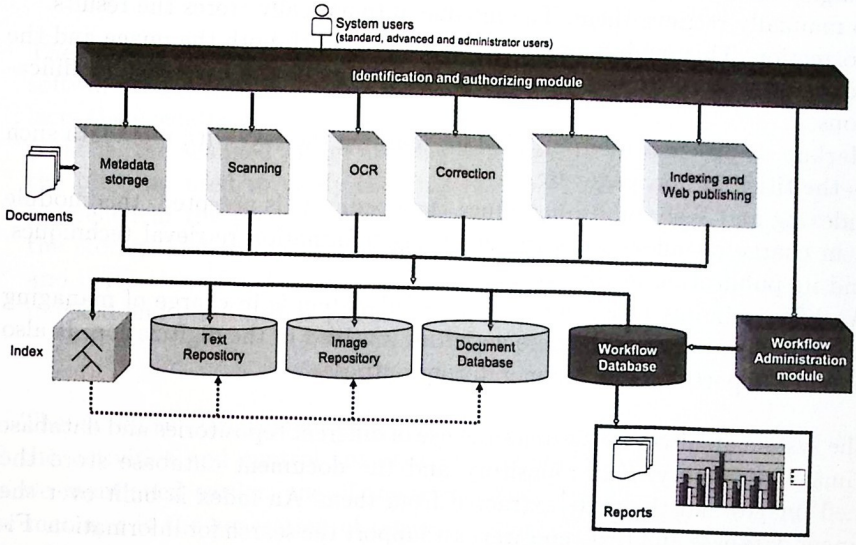

Fig. 1. System architecture

Figure 1 shows the overall system architecture. When defining it, we followed the recommendations of the Workflow Reference Model [8], a commonly accepted framework for the design and development of workflow management systems. Intended to accommodate the variety of implementation techniques and operational environments which characterize this technology. Thus, although we used this architecture for the implementation of a specific system, it can be used in other environments and situations.

As we can see in Figure 1, the authentication and authorizing module is in charge of the authentication of the workers who want to access to the system. Each user has a system role depending on the tasks he/she is going to work on. In terms of this system role, the authorizing module only provides the user with access to the needed features. The system architecture is composed of a module for each activity carried out during the repository creation.

- Metadata storage. This subsystem is in charge of the introduction and storage of the metadata for each document (title, author, year, source, etc.), following any desired format, as Dublin Core [10] or MARK [11]. This task is performed by the advanced users of the system, therefore only them have access to this module.
- Scanning. This system provides access to the scanning hardware and software, and it is the responsible of managing the specification of the scanning parameters for each document (for example, options like scanning two pages at the same time, landscape orientation, resolution, number of colors, etc.).
- OCR. It provides access to the OCR software, and obtains the scanned images needed as the input of this activity, therefore it is not necessary to manually retrieve them. The module automatically stores the results.
- Correction. This module provides the reviewer with both the image and the extracted text to carry out the correction to make the necessary modifications.
- Markup. It provides the tools used for marking the text with metadata such as the title, author, page, etc.
- Indexing and Web publishing. Once the document is accepted, this module is in charge of indexing its content using information retrieval techniques, and its publication in the Web.
- Workflow administration module. This subsystem is in charge of managing the workflow between all these activities involved in the digitization. It also provides reporting tools for monitoring purposes.

The system architecture assumes the use of different repositories and databases. The image repository, text repository and the document database store the scanned images and the texts extracted from them. An index is built over the document database and text repository to support the search for information. Finally, the workflow database stores the information about the digitization chain, with the lists of tasks, the state of each document, etc.

5 DigiFlow: A Tool for Document Repository Building

The architecture presented in the previous section was applied in the implementation of DigiFlow, a workflow management system supporting the digitization process of digital libraries. This system was used in the case study presented in next section. Thus, in this section we briefly describe some issues regarding to the system implementation.

DigiFlow provides an integrated environment where all the tasks can be executed. The idea is that this application provides the user with all the tools needed to carry out each task, without being necessary to use other software applications or manually manage the results of each task. Once the user credentials are introduced and validated in the system, the application shows the list of pending tasks. Once he/she selects the execution of any of them, the system

retrieves the necessary resources and opens the corresponding applications. For example:

- When the user selects a scan activity, the system shows the specification of the task to the user, indicating the document identifier, the page numbers to scan and the scanner options. All this information is shown within the same application. When the user presses the "Scan" button, the document is scanned with the specified options and once the user confirms the validity of the obtained result, the system stores the resulting image and creates the corresponding OCR task. Notice that the user does not have to manually manage any of this products.

- The same approach is followed by the execution of the correction task. The software used for this activity is OmniPage Pro 12 [12]. When the user selects the execution of this task, DigiFlow retrieves the inputs for the review (that is, the scanned image and the text extracted with OCR) and opens the program used to verify that the text corresponds with the image. As happened in the previous example, the user does not have to care about the storage and retrieval of the intermediate documents. When the review and correction of a document is completed, the document is automatically indexed and published in the Web of the digital library.

The workflow management subsystem provides all the tools needed to define all the activities and control the results obtained in each of those activities. As we mentioned earlier, one of the most useful tools the system provides to the manager is the generation of reports about the digitization process, which allow him/her to monitor and control the progress of the project and to identify possible problems. DigiFlow offers three types of reports to the administrator:

- Opened document report. This report informs the administrator of the current state of each document, the tasks to be done and the results obtained in each of them.

- Activity report. This report shows the administrator statistics about each activity of the digitization process. This allows the administrator to have an idea of the effort devoted to each activity and to identify possible bottle necks.

- Users report. This report summarizes the dedication of each worker to each task, such as the work load for each of them and the degree of completion of the assigned tasks. This allows the administrator to assign new activities to the free workers.

6 Case Study: A Digital Library for Ancient Documents

In this section we present the results obtained in the application of DigiFlow to build a digital library of an institution responsible of the care of an important collection of old documents.

6.1 The institution - *Royal Galician Academy*

Royal Galician Academy (RAG) is an institution devoted to the conservation and promotion of the cultural heritage and the language of Galicia, a region located in the northwest of Spain [13]. One of the main characteristics of this region is its rich culture, its own language, and a rich literature. RAG is the responsible of the maintenance and conservation of an archive that stores a big collection of old documents. As an initiative for the promotion of this collection, this institution decided to build a digital library which is compounded of journals of the 19^{th} century. These journals are a patrimony of great value, being many of them unique and showing the situation of Galicia at that time. The publication of this documents in a digital library facilitates their conservation and makes them available to any interested researcher and to the public.

The whole collection contains 23,000 documents. Due to their antiquity and their state of conservation, the digitization of this documents was really difficult. An example of these documents is shown in Figure 2, and all the digitized documents can be accessed in [14]. The size of the collection and the deep knowledge needed to correct the mistakes that appear in the digitization also contributed to the difficulty in building the document repository.

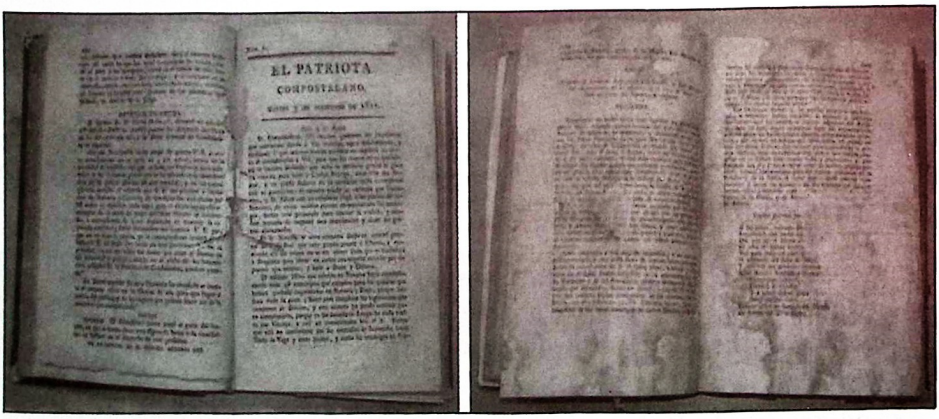

Fig. 2. Images from "El Patriota Compostelano" (1810)

6.2 Building the digital library

The building of the document repository was carried out by a collaboration between the Databases Laboratory (our research group) and the Laboratory of Linguistic Technology of University of A Coruña. A group of twenty granted students from the Laboratory of Linguistic Technology worked in the digitization of documents. Due to their deep knowledge of Spanish and Galician language and of this kind of documents, they are skilled enough to successfully carry out this work.

The hours spent in each one of the main workflow activities are shown in Table 1, as well as the number of pages processed in that period. This information can be used to monitor and take under control such a costly process, and plays an important role as estimation data for future projects of the same characteristics that are planned to be executed. For example, in this information we can see quantitatively the high cost of the correction tasks in comparison with the others when working with old documents in a bad state of conservation.

ACTIVITIES	PAGES	HOURS	PAGES/HOUR
Metadata storage	23,000	235.99	97.46
Scanning	23,000	442.17	52.02
OCR	22,635	658.37	34.38
Correction	21,077	7,611.49	2.77

Table 1. State of the work.

7 Conclusions and future work

The creation of a document repository is not a simple process. It requires the coordination of people and tools to carry out every activity that is part of the process. These activities include digitization of documents, optical character recognition, results correction and indexing to perform search by content. For all these process to be correctly and efficiently made, it is necessary the use of support tools that facilitate the work of each participant and ensure the quality of the obtained results.

The proposed workflow strategies and system architecture support the control and coordination of people and tasks involved in the digitization process. The use of this architecture automates the completion of prone to error activities and optimizes the performance of the digitization process and the quality of the obtained results. This architecture was applied to the design and development of DigiFlow, a collaborative workflow management system designed to create document repositories. This system was built as a desktop application which provides an integrated environment for the execution of all the tasks.

DigiFlow was successfully applied in the building to a digital library for "Royal Galician Academy", an institution devoted to the conservation and promotion of the Galician culture and language. This digital library involved the

digitization of a set of 23,000 documents of the 19^{th} century in a bad state of conservation. In this project, the proposed architecture showed its benefits automatizing all the digitization chain. In this paper, we present this experience with quantitative results of the effort devoted to each task of the project. Although the system was used in a specific case, the proposed architecture is general enough to be applied in any project of these characteristics.

Our work line still maintains opened some questions that will be addressed in the future. First, the developed system is going to be applied to new projects of similar characteristics involving the digitization of hundreds of thousands of pages. In addition, we are working on different implementations of the activities considered to optimize the performance of the overall process. We are also working to provide the system with new compression and indexing tools necessary for really big digital libraries.

References

1. Ross, S., Hedstrom, M.: Preservation research and sustainable digital libraries. International Journal on Digital Libraries **5**(4) (2005) 317–324
2. Borgman, C.L.: Challenges in building digital libraries for the 21st century. In: 5th International Conference on Asian Digital Libraries, ICADL 2002, Singapore, December 11-14, 2002. Proceedings. (2002) 1–13
3. Chang, N., Hopkinson, A.: Reskilling staff for digital libraries. In: Digital Libraries: Achievements, Challenges and Opportunities. Volume 4312 of Lecture Notes in Computer Science., Springer (2006) 531–532
4. McCray, A.T., Gallagher, M.E.: Principles for digital library development. Communications of the ACM 44(4) (2001) 49–54
5. Witten, I.H., Bainbridge, D., Boddie, S.J.: Power to the people: End-user building of digital library collections. In: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries JCDL '01. (2001) 94 – 103
6. Bainbridge, D., Thompson, J., Witten, I.H.: Assembling and enriching digital library collections. In: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries. (2003) 323 – 334
7. Hollingsworth, D.: Workflow management coalition - the workflow reference model. Technical report, Workflow Management Coalition (1995)
8. van der Aalst, W., van Hee, K.: Workflow management : Models, methods, and systems. (2002)
9. Fischer, L.: Workflow handbook 2003. Future Strategies Inc., USA (2003)
10. Hillmann, D.: Using dublin core. Technical report, Dublin Core Metadata Initiative (2005)
11. Furrie, B.: Understanding marc - bibliographic machine readable cataloging. Technical report, Library of Congress - Network development and MARC standards office (2003)
12. Scansoft: Scansoft omnipage pro web page. (Web page) http://www.scansoft.com/omnipage.
13. RAG: Real academia galega web page. (Web page) http://www.realacademiagalega.org.
14. RAG: Real academia galega digital library. (Web page) http://www.realacademiagalega.org/Hemeroteca.