A Tool for Workflow Management in the Composition of Multimedia Databases from Preexistent Documents

Eduardo A. do Valle Jr.[1], Arnaldo de A. Araújo[1],
Fernanda M. Vieira[1], Carla C. P. da Costa[2]

[1]NPDI/DCC – Universidade Federal de Minas Gerais
Caixa Postal 702 – CEP 30123-970 – Belo Horizonte – MG – Brazil

[2]Arquivo Público Mineiro
Av. João Pinheiro, 372 – CEP 30130-180 – Belo Horizonte – MG – Brazil

{edujr, arnaldo, nanda}@dcc.ufmg.br, carlac@apm.mg.gov.br

Abstract. *When reformatting a large collection of documents in order to compose a multimedia database, a project manager faces two considerable challenges. One is to ensure the digitization process follows precise and consistent standards through the project lifecycle. The other is to make the acquisition of managerial metadata as straightforward as possible, and automatic whenever possible. Instructing the operator in every step of the acquisition process, and concomitantly registering the relevant metadata, the AdLabore tool provides a simple answer to both those questions.*

1. Introduction

Many multimedia databases are created by reformatting preexistent conventional collections. The process of digitization can be applied to many types of artifacts - audio tapes, *Long Play* records, negatives and slides films, pictures, celluloid movies, analog video tapes – often in large batches. The custodian may be interested in different advantages of digital technology: making access easier and broader, using the digital form as auxiliary for preservation, gaining the flexibility that digital data experiences, etc.

In order to get these advantages, however, the conversion process must be carefully managed and executed. A careless digitization process raises all sorts of problems varying from missing some of the contents, to creating longevity issues due to poor standardization. Unfortunately, merely training the operators before the digitization starts and instructing them to type the necessary metadata in a separate activity is highly inefficient. The digitization task is at the same time delicate and repetitive, and thus very prone to human error. Adding the metadata creation as a parallel but independent process adds another opportunity for inaccuracy and forgetfulness.

Obviously, when the digitization process can be fully automated – by using document feeding, batch processing, automatic signal filtering/enhancing, and automatic feature extraction – those disadvantages can be avoided. But this complete automation can almost never be obtained. Fragile documents, hard-to-read data, manuscripts and bound volumes are today impossible to digitize and index without manual intervention.

In order to coordinate the complex activity of digitization, keeping the operator in the limits of the process standards, registering all the significant information about the tasks execution and optimizing the alternation of manual and automatic operations we can use a workflow management tool. Workflow management systems store a representation of the processes the user wants to enact, and coordinates the *events* (i.e., the processes enactment), automating some steps, taking care of the bookkeeping and giving the user precise instructions [CICHOCKI 98].

Libraries, archives and other custodian institutions are great potential users of these tools, because of the size, value and sophistication of their collections. The recent trend of digital libraries adds momentum to this necessity: *As libraries convert materials into digital format, the need for efficient workflow management tools will increase. The Potential of digital libraries is great, but the substantial effort, from a workflow perspective, is equally, if not more, daunting* [CHOUDHURY 2000]. These users are especially interested, while reformatting their collections, in the following aspects:

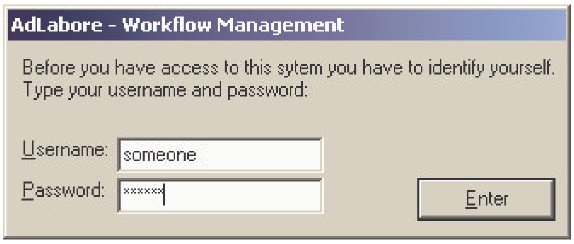
1. The preservation of the identity of the digitized material, i.e., warranting that there will be a way of mapping the digital files to the physical artifacts;
2. The adherence to standards in the digitization process. In some cases those can be as lax as just defining a general file format and a small set of constraints (e.g., maximum acceptable bandwidth); In other cases the standards can be complex and strict, prescribing precise framing, special lighting, etc.
3. The acquisition of some information about the document being digitized (intellectual metadata) and about the digitization process itself (managerial metadata). This can be very important especially for preservation purposes.

In this paper we present *AdLabore*, a workflow management tool of simple conception, but very powerful in these situations. Its main strength is blending the process of instructing the operator and collecting the necessary metadata in one single flow of interaction. The operator decisions are made straightforward, but he/she is kept alert to the process because the system is constantly requiring feedback.

The available processes can be fully customized by the project manager, and AdLabore can interact with other applications. AdLabore is written in *Microsoft Visual Basic*, for all the 32-bits *Microsoft Windows* environments. It is available to the public, as *freeware*, and its sources are available through a *copyleft* license.

2. Guiding the Operator

The user interface is built to be as simple and effective as possible. The operator starts AdLabore by identifying him/herself, as shown in Figure 1.

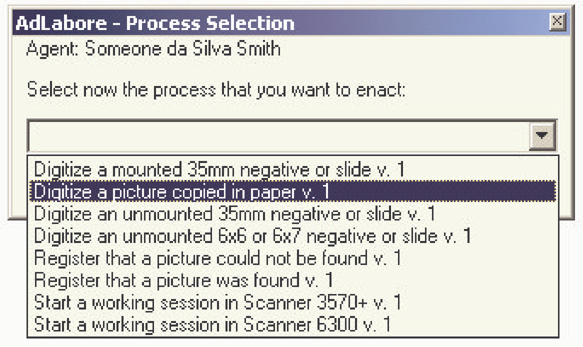


Before you have access to this sytem you have to identify yourself.
Type your username and password:

Username: someone
Password: ******

Figure 1. Agent identification dialog box.

Then he/she can choose what process to enact, as shown in Figure 2.



AdLabore - Process Selection

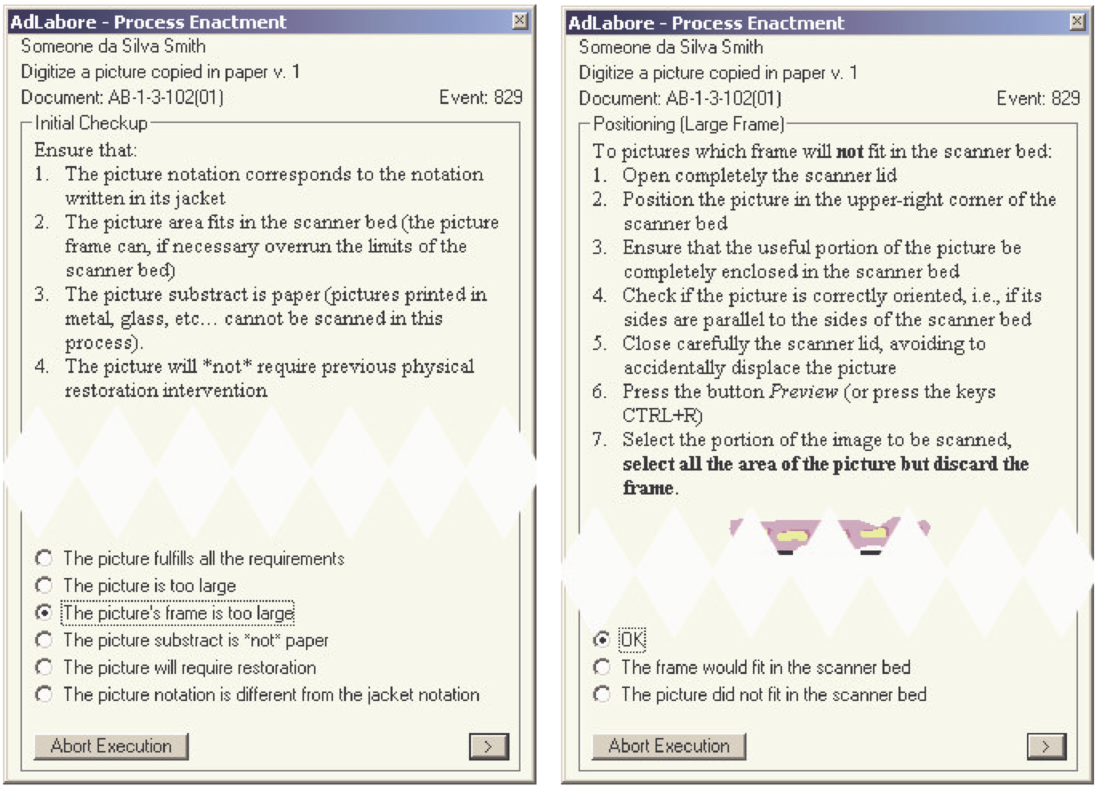
Agent: Someone da Silva Smith

Select now the process that you want to enact:

Digitize a mounted 35mm negative or slide v. 1
Digitize a picture copied in paper v. 1
Digitize an unmounted 35mm negative or slide v. 1
Digitize an unmounted 6x6 or 6x7 negative or slide v. 1
Register that a picture could not be found v. 1
Register that a picture was found v. 1
Start a working session in Scanner 3570+ v. 1
Start a working session in Scanner 6300 v. 1

Figure 2. Process selection dialogue.

A process enactment is a dialogue with the application, composed of many steps. The operator is instructed in how to perform some task, and then is expected to provide some data. Being a multimedia application itself, AdLabore can instruct the operator in many forms, by using a composition of text, audio, video and still images. The expected answer may be a multiple-choice question or a short text entry (see Figure 3).



AdLabore - Process Enactment

Someone da Silva Smith
Digitize a picture copied in paper v. 1
Document: AB-1-3-102(01) Event: 829

Initial Checkup

Ensure that:

1. The picture notation corresponds to the notation written in its jacket
2. The picture area fits in the scanner bed (the picture frame can, if necessary overrun the limits of the scanner bed)
3. The picture substract is paper (pictures printed in metal, glass, etc... cannot be scanned in this process).
4. The picture will *not* require previous physical restoration intervention

○ The picture fulfills all the requirements
○ The picture is too large
● The picture's frame is too large
○ The picture substract is *not* paper
○ The picture will require restoration
○ The picture notation is different from the jacket notation

AdLabore - Process Enactment

Someone da Silva Smith
Digitize a picture copied in paper v. 1
Document: AB-1-3-102(01) Event: 829

Positioning (Large Frame)

To pictures which frame will **not** fit in the scanner bed:

1. Open completely the scanner lid
2. Position the picture in the upper-right corner of the scanner bed
3. Ensure that the useful portion of the picture be completely enclosed in the scanner bed
4. Check if the picture is correctly oriented, i.e., if its sides are parallel to the sides of the scanner bed
5. Close carefully the scanner lid, avoiding to accidentally displace the picture
6. Press the button *Preview* (or press the keys CTRL+R)
7. Select the portion of the image to be scanned, **select all the area of the picture but discard the frame**.

● OK
○ The frame would fit in the scanner bed
○ The picture did not fit in the scanner bed

Figure 3. While enacting a process, the operator is oriented in how to perform every step and is expected to provide a piece of information.

2. Collecting Metadata

All information about the material that is being reformatted, besides the digitized files themselves, compose what is called *metadata*. Metadata capture important information about the data, such as *intellectual information* (author, creation date, title...), *rights* (copyright, access rights...), *behavior* (hyperlinks, sequencing, ...) *format* (resolution, bit depth, file format...) and *management* (who digitized, digitization date...) [HOPKINS 98].

In AdLabore metadata are generated partially automatically by the process (e.g., the process can keep track of who is enacting it, the start time, end time, etc...) and partially based on the answers of the operator. All metadata are stored in a symbol table for every event and can be later retrieved.

3. Composing the Process

When composing the workflow, one can choose among many paradigms for describing the processes. One approach is to use artifact-based modeling, where the documents (both physical and digital) are the main concern of the system. Another is activity-based modeling, where the actions performed are the main description and tracking unit [CICHOCKI 98]. Because AdLabore is dedicated mainly to document reformatting, its paradigm is highly artifact-based, but since activity-based modeling is much easier to understand it uses a combination of the two approaches.

There is a high degree of flexibility in how new processes can be composed. Each process can be chosen to be **documental** (more artifact-based) or **non-documental** (more activity-based). Then they are described as an ordinate series of steps, that can be composed of *scriptlets*, instructions with text and multimedia, questions that the operator must answer before proceeding, etc... The *scriptlets* allow some extent of automation, especially in the generation of metadata, performing simple computations, and interacting with other applications.

Of course, composing good processes is critical to the success of the reformatting effort. The best processes are those that divide a complex activity in simple steps, avoid subjective decisions, are clear and precise but require constant interaction and alertness of the operator, in order to avoid boredom and consequent failures.

4. Acknowledgements

The authors wish to thank CNPq, CAPES and the SIAM Project DCC/PRONEX for supporting this work

Eduardo A. do Valle Jr. is sponsored by a scholarship provided by CNPq / Brazil.

References

CHOUDHURY, S. et al. (2000), "Digital Workflow Management: The Lester S. Levy Digitized Collection of Sheet Music" in First Monday, volume 5, number 6. http://firstmonday.org/issues/issue5_6/choudhury/index.html

CICHOCKI, A. et al., Workflow and Process Automation: Concepts and Technology. Kluwer Academic Publishers, Boston, 1998.

HOPKINS, J. (1998) “What is metadata?”. Accessed in June/2002.
http://ublib.buffalo.edu/libraries/units/cts/preservation/metadata.html.

VIEIRA, F., VALLE, E. et al. (2002) “Um Sistema de Workflow para Documentos Históricos”. In: Anais da X Semana de Iniciação Científica, Universidade Federal de Minas Gerais. Brazil.