Optimization in Historical Digital Library Workflows

Optimização de Fluxos de Trabalho em Bibliotecas Digitais Históricas

João Manuel Fernandes Cardoso
Instituto Superior Técnico, Campus Taguspark
Av. Prof. Doutor Aníbal Cavaco Silva
2744-016 Porto Salvo
joao.m.f.cardoso@ist.utl.pt

ABSTRACT

Humankind have always felt the need to record its endeavours. Libraries are a natural consequence of that everlasting desire. They are but locations where information is stored, preserved and made accessible for those who seek it. With the advent of the information age, digitization of physical records became common practice, and thus the first "digital libraries" were born.

This article focuses on optimizing the workflows from an historical digital library Hemeroteca Municipal de Lisboa. It presents a study on the business processes at Hemeroteca's digitizing and image service, and the workflows they generate. Finally it describes the implementation of optimizations to those business processes and analyses the results from that implementation.

Categories and Subject Descriptors

[**Applied computing**]: Digital libraries and archives
; [**Information systems**]: Process control systems
; [**Information systems**]: Multimedia content creation
; [**Information systems**]: Document structure

Keywords

Digital Library, Digitization, Metadata, Digital Publishing, Hemeroteca, Business Process, Workflow

1. INTRODUCTION

Humankind have always felt the need to record its endeavours. Cave paintings were the first form of recording known to man, with the earliest records dating back to over 40,000 years. The first libraries were used to record the earliest form of writing: clay tablets in cuneiform script dating back to 2600 BC. Undoubtedly the most notorious of all the great libraries of the ancient world was the Royal Library of Alexandria in Egypt, around 300 BC.

The concept of library, a location where collections of records were preserved, stored and made accessible, didn't experience radical change until the arrival of the information age. Only then advances in computer science allowed for libraries to start working with digital representations of its physical records. Thus, the concept of "digital library" was born, having only appeared in the early 1990s with the advent of the Internet.

This article focuses on optimizing the workflows from an historical digital library, the Hemeroteca Municipal de Lisboa (Hemeroteca)[1]. This will imply studying the business processes at Hemeroteca's digitizing and image service, and the workflows they generate. Only fully understanding and optimizing the business processes can those optimizations be reflected in the workflows they generate.

The study of the business processes at the digitizing and image service at Hemeroteca was made by resorting to two distinct sources. Firstly, a general technical analysis of the domain was carried out, through the investigation of the main topics and related state of the art. Secondly, a pragmatic raising of requirements was carried out, based on the analysis of the actual case at the Hemeroteca and on interviews to the actors that are currently in charge and operating the business processes at the digitizing and image service. It was only by fully comprehending the actors needs and faults that any solutions aiming to optimize their business processes could be proposed and executed.

These solutions ranged from software applications which aimed to either optimize tasks which are manually performed or completely eliminate the human factor by automating them. However it is important to mention that not all solutions were self evident. On occasion constraints originated from the context of the problem, had an impact on the development and design of the solutions that are presented within this article. Additionally training was also provided for the existing staff members in order for them to improve their performance while performing tasks.

Overall this article will show that the objectives that were proposed during the project phase of this thesis were accomplished. Not only that, but their success led to the improvement of both the efficiency and effectiveness of a digital library, thus proving that masters thesis can and should have a real world impact.

1.1 Motivation and objectives

The generic objective of this work is to propose solutions to optimize workflows which result from business processes for digitization and publishing in an existing cultural her-

[1] Hemeroteca Municipal de Lisboa website: http://hemerotecadigital.cm-lisboa.pt/

itage digital library, mostly comprised of newspapers and magazines whose intellectual property rights have expired, i.e., have fallen into public domain.

The concrete focus and motivation is as stated above the Hemeroteca Municipal de Lisboa, specifically their digitizing and image service. The optimizations to be proposed, implemented and validated must contribute so that the tasks that make up their business processes can be performed more efficiently (using less resources and time) and more effectively (improving the quality of the final product and providing in the end a better experience to the final users when interacting with the digital library).

1.2 Article Structure

This article is divided into four different sections. Section 1 presents the introduction, motivation and objectives for this thesis article. Througout section 2 an analysis on the problem that was tackled in this thesis project is made, and the results of the assessment made on the Hemeroteca are detailed. section 3 on the other hand describes the solutions and their designs to the issues related in section 2. Finally, section 4 finishes this article by drawing conclusions on the execution of this thesis project, and proposes future changes to the solutions that might improve on what has been achieved.

2. PROBLEM ANALYSIS

This section focuses on both contextualizing and analysing the problem that is the aim of this dissertation. Therefore a closer look is taken at the procedures and structure of the Hemeroteca Municipal de Lisboa (Hemeroteca) for the optimization of its business processes workflow is this thesis project's focus case.

This section is divided into three sections. Section 2.1 presents a contextualization of the focus case of this dissertation. Section 2.2 goes through the business processes in play at the Hemeroteca. And finally section 2.3 displays the results of the analysis made on the business processes and identifies the problems with a specific selection of subprocesses.

2.1 Hemeroteca

The Hemeroteca is under the supervision of Câmara Municipal de Lisboa (CML). Its objective is to create and maintain a cultural heritage digital library, mostly comprised of newspapers and magazines whose intellectual property rights have expired, i.e., have fallen into public domain. In its own domain, Hemeroteca is Lisbon's second largest digital library, holding over half a million records, with a production between 140 to 150 thousand new records per year.

The Hemeroteca is expected to change its facilities in 2014, however it has been based on the palace of the counts of Tomar, which is located at Bairro Alto in Lisbon, since the year 1973. Currently the facility is over-capacitated, thus justifying the upcoming move.

Hemeroteca's digitization and image service has as a main goal the digitalizing and publishing of works whose rights have already fallen into public domain. There are four people currently attached to this service. Anabela Ferreira, João Oliveira, Joaquina Cunha and Paula Cardoso. All four staff members have years of experience working in the field of digital libraries, however only three have had specific training to do so. Both Joaquina Cunha and Paula Paula have a technical course in articles and libraries, and João Oliveira holds a graduate course in information and documentation sciences, specialized in libraries.

The service's workload and responsibility is split between all the four members of staff, however due to the nature of his training João Oliveira occupies a *de facto* management position within the service. With Alvaro de Matos overseeing the service as its coordinator.

In terms of resources, the digitization and image service has four flatbed scanners, with plans on acquiring a book scanner after the move to the new facilities in 2014. It has five computers which have Intel Core 2 Duo processors, with RAM capacities ranging from 900MB to 3GB, all of which are equipped with Microsoft's Windows XP Professional operating system.

As for the software tools in use by the staff of the digitization and image service, eight software tools should be considered:

- **Microsoft FastStone Image Viewer**[2] An image browser, organizer, converter and editor designed for Microsoft Windows by FastStone Soft and provided free of charge for non-commercial use;

- **PAPAIA**[1, p. 3] A software tool which was originally developed to be used at the Biblioteca Nacional Digital (BND). PAPAIA processes batches of images, and can preform actions such as: renaming images, editing the TIFF headers and registering structural metadata;

- **Adobe Acrobat** Hemeroteca holds one Adobe Acrobat[3] software licence. Adobe Acrobat is a software application developed by Adobe Systems to view, create, manipulate, print and manage files in Portable Document Format (PDF);

- **JPEGToPDF**[4] A freeware software application used to convert Joint Photographic Experts Group (JPEG) image files to PDF, that does not require Adobe Acrobat or Acrobat Reader ;

- **BecyPDFMetaEdit**[5] A freeware software application which loads PDFs and allows editing of its descriptive metadata, i.e., author, title, subject and keywords of the document;

- **ContentE**[6] [2] A software application developed by Gilberto Pedrosa. It produces master copies for preservation, copies for access, structural descriptions in Metadata Encoding and Transmission Standard (METS), and also indexes. The master copies are organized within a folder structure, which has a folder for each Multipurpose Internet Mail Extensions (MIME)[4] (e.g., Tagged Image File Format (TIFF), JPEG, Portable

[2] Microsoft FastStone Image Viewer: `http://www.faststone.org/` Retrieved 2013-09-30

[3] Adobe Acrobat website: `http://www.adobe.com/pt/products/acrobat.html` Retrieved 2013-09-30

[4] JPEGToPDF website: `http://www.jpegtopdf.com/` Retrieved 2013-09-30

[5] BecyPDFMetaEdit website: `http://www.becyhome.de/becypdfmetaedit/description_eng.htm` Retrived 2013-09-30

[6] ContentE website at the BND: `http://purl.pt/index/geral/PT/infoProfContentE.html` Retrived 2013-09-30

Network Graphics (PNG), Graphics Interchange Format (GIF), PDF or Text File (TXT));

- **Microsoft FrontPage** An Hypertext Markup Language (HTML) editor and web site administration tool which was developed by Microsoft and was distributed with Microsoft Office from 1997 to 2003. However it has since been discontinued;

- **WinSCP (Windows Secure CoPy)**[7] A free and open-source SFTP, SCP and FTP client for Microsoft Windows. It offers secure file transfer between a local and a remote computer as well as a basic file manager and file synchronization functionality.

As for Information Technology (IT) support, it is provided by the Departamento de Modernização e Sistemas de Informação (DMSI), which is inserted within the Divisão de Administração de Sistemas e Infrastructuras (DASI) of the CML. This delegation of responsibility means that none of the four members of staff currently working at Hemeroteca's digitization and image service has any experience in managing information systems. This is an important constraint to the design presented in section 3.

2.2 Business Processes

Hemeroteca has two business processes. The Process Digitizing and Publish (P1) has as overall objective the publishing of a digitized work, and it is comprised by four tasks that can be perceived as a collapsed sub-processes, as can be seen in figure 1. A brief description of each sub-process is given in table 1. However since the stakeholder decided that the Sub-Process Digitizing (P1.1) should not be considered for optimization, only the remaining three sub-processes will be detailed in sections 2.2.1 through 2.2.3.

The Process Metadata Sharing (P2) has as main goal updating a catalogue hierarchy of which Hemeroteca's catalogue sits on the bottom. As it was decided by the stakeholder not to tackle this business process, the tasks that make up this process will not be described in this article.

2.2.1 Sub-Process File Name Normalization (P1.2)

The Sub-Process File Name Normalization (P1.2) (see figure 1) consists on the naming of the images according to a determined naming convention, which is in use at the BND[3]. This sub-process serves two purposes: The identification of the image through a unique identifier, and the display of the image's technical features in its name. This procedure is accomplished through the use of a software tool named PAPAIA (see section 2.1). Using PAPAIA the staff at the digitization and image service processes the JPEG image files that originated in the digitization sub-process. The end product is a renamed set of JPEG image files.

2.2.2 Sub-Process Metadata Editing (P1.3)

The Sub-Process Metadata Editing (P1.3) (see figure 2) starts by picking up on the products from the file name normalization sub-process, and proceeds to create both a PDF file that aggregates all the JPEG image files and structural metadata describing how the JPEG image files should be organized to reproduce the original publication's structure.

Hemeroteca's digital objects are attainable to users as PDFs, and JPEG image files (the latter inserted within HTML pages). This implies that the creation of PDF files must be a necessary task within the P1.3. However only a single Adobe Acrobat licence is held at the digitization and image service, this implies that the remainder of the staff must use a freeware software application (JPEGToPDF, see section 2.1) to create PDF files.

The task that follows is the enrichment of the PDF files with descriptive metadata. Again if the staff member that is preforming the P1.3 does not hold an Adobe Acrobat licence, he will be forced to use a freeware application (BecyPDFMetaEdit, see section 2.1) to perform the task.

The structural metadata editing task is performed with the assistance of a software tool called ContentE (see section 2.1), which also produces access copies as Extensible HyperText Markup Language (XHTML) files.

2.2.3 Sub-Process Publishing (P1.4)

The Sub-Process Publishing (P1.4) (see figure 3) consists on the creation of HTML files which will then be published at a external server which is managed by the DMSI-DASI. The whole sub-process relies heavily on manual tasks being performed by staff members of the digitizing and image service.

The first task of the sub-process depends on whether the work to which the new publication belongs exists at Hemeroteca's work index. The work index is a web page containing the titles of all the works that have been digitized to date. If the work does not exist, then it must be added to the work index, and that involves creating a new work page. If the work exists in the work index, then the new publication is simply added to the existing work page. Both these tasks are performed using FrontPage (see section 2.1), which means manually editing HTML files.

Once the necessary work related HTML files have been created, the next task is to compile a list of all the authors which collaborated on the publication. This will be used to check whether all the listed authors exist at the author index. If any of them does not exist they must first be added to the author index and a new author page must be created for that author. Regardless of the author existence in the author index, the work must be added to each individual author page. Again, these task are performed using FrontPage.

The final task is to send all newly generated or edited HTML files to the external server which, as mentioned, is managed by the DMSI-DASI. This is done by using a safe connection through the SFTP, SCP and FTP client for Microsoft Windows WinSCP (see section 2.1).

2.3 Consolidated Analysis

This section analyses the problems that are to be tackled throughout this thesis project's execution, (which is detailed in section 3).

As stated in section 2.2, the P1 at Hemeroteca comprised four tasks that can be perceived as a collapsed sub-process. Since the stakeholder defined the P1.1 should not be considered for optimization. Only the remainder three sub-processes were analysed, quickly becoming apparent that all of them could suffer improvements that would not only simplify them, but also lead them to be performed more efficient and effectively.

[7] WinSCP website: `http://winscp.net/eng/index.php` Retrieved 2013-09-30

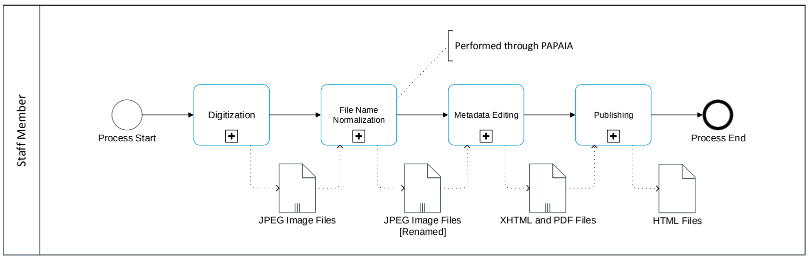


Figure 1: The Process Digitizing and Publish (P1)

Activity Name	Summarized Description
Digitization	Creation of a digital images by digitizing an original publication
File name normalization	Renaming the image files according to a specific naming convention
Metadata editing	Creating PDF files, and editing both the descriptive and structural medatada of the digital publication
Publishing	Publishing the digital publication at Hemereoteca's website

Table 1: First business process activities

The P1.2 presented a challenge because it was based on an outdated software application (see section 2.2.1). PAPAIA, as described in section 2.1, was originally designed for the BND. However it was never able to display its full potential when in use by the digitizing and image services at Hemeroteca. Features such as editing TIFF headers or registering structural metadata were either never performed, or were alternatively completed through the use of different software applications. PAPAIA was solely used for renaming batches of image files, and its interface proved to be so complex that none of the four staff members of the digitizing and image services fully understood how to take advantage of the renaming abilities that PAPAIA offered. All image files were thus renamed with the same scheme, the only variable being its order number.

The P1.3 was riddled with redundant tasks, particularly in what referred to the descriptive metadata editing. There were two different software applications being used to both create and add descriptive metadata to PDF files. In terms of the structural metadata addition, the version of the software application ContentE was outdated (The current version is 3.9, whilst the version used at the digitizing and image service is 1.6). The staff at the digitizing and image service had also not been specifically trained for its use, and therefore could not take full advantage of the application's potential.

As can be seen in section 2.2.3, the P1.4 relies heavily on manual procedures and has little automation. This means that the staff members at the digitizing and image services spend a heavy portion of their time editing HTML files, when they could be spending it processing more works. Additionally the manual editing of HTML files leads to syntax errors, which hamper the consistency of data publishing.

3. DESIGN AND SOLUTION

This section focuses on presenting solutions for the problems stated in section 2.3.

The section is divided into three sections. Section 3.1 goes through the solution to problems related to the file name normalization sub-process (see section 2.2.1), and presents the design to the software application that was developed to solve them. Section 3.2 focuses on the measures that were taken to solve the issues related to the metadata editing sub-process (see section 2.2.2). Finally section 3.3 analyses the solution presented for the publishing sub-process (see section 2.2.3), and describes the design of the software application that was developed for that intent.

3.1 File Name Normalization

As mentioned in section 2.3 there were three main issues that needed to be tackled in order to optimize the P1.2. These issues were:

- Image files needed to be renamed according to the naming convention used at the BND.
- PAPAIA was an outdated software application, that had been designed specifically for the BND;
- The staff at the digitizing and image service should be able to understand and take advantage of all of the software application's features.

Having these issues in mind, a decision was made to scrap PAPAIA and replace it for an application that would not only fit the digitizing and image service needs more efficiently, but also allow its staff to take full advantage of its features.

The software application that was to replace PAPAIA was called Carica. The name Carica is a wordplay based on the fact that Carica is a genus of flowering plants in the family Caricaceae, which includes the papaya. PAPAIA's substitution would not imply any changes to the P1.2 itself, for changes would only occur in terms of how the task was performed and not on the definition of the task itself.

Carica was to replicate PAPAIA's image file renaming feature. However it upgraded it by going a step further and allowing users to create, edit and apply a renaming schemas to a batch of image files. It also allows users to create, edit and apply a set of renaming schemas to a batch of image files (see figure 4). There are two main concepts behind Carica. These are the concept of schema and work, which are described in table 2.

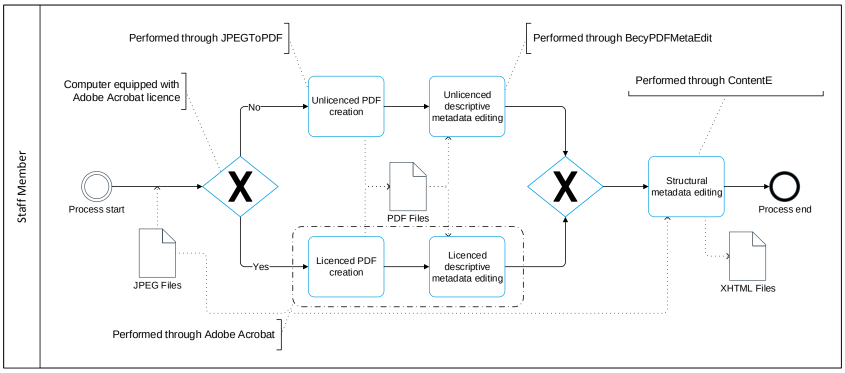

Staff Member
Performed through JPEGToPDF
Performed through BecyPDFMetaEdit
Computer equipped with Adobe Acrobat licence
Performed through ContentE
No
Unlicenced PDF creation
Unlicenced descriptive metadata editing
Process start
X
PDF Files
X
Structural metadata editing
Process end
JPEG Files
Yes
Licenced PDF creation
Licenced descriptive metadata editing
XHTML Files
Performed through Adobe Acrobat

Figure 2: The Sub-Process Metadata Editing (P1.3)

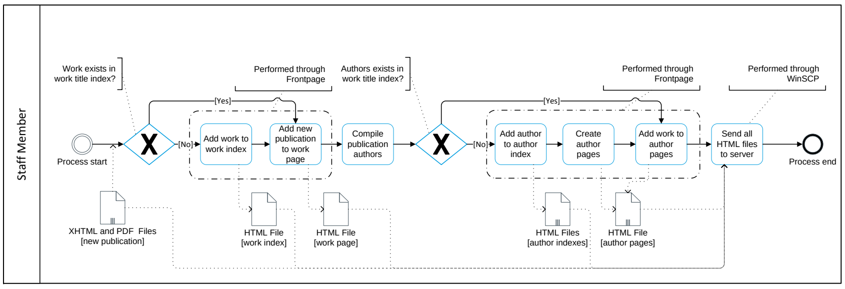
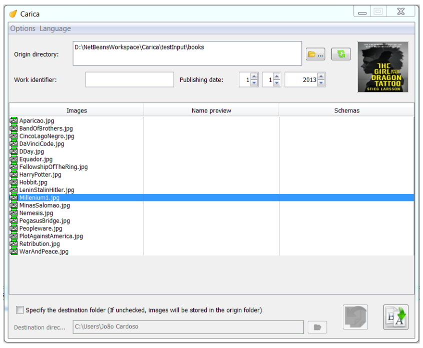

Staff Member
Work exists in work title index?
Performed through Frontpage
Authors exists in work title index?
Performed through Frontpage
Performed through WinSCP
[Yes]
[Yes]
Process start
X
[No]
Add work to work index
Add new publication to work page
Compile publication authors
X
[No]
Add author to author index
Create author pages
Add work to author pages
Send all HTML files to server
Process end
XHTML and PDF Files [new publication]
HTML File [work index]
HTML File [work page]
HTML Files [author indexes]
HTML File [author pages]

Figure 3: The Sub-Process Publishing (P1.4)



Carica

Options Language

Origin directory: D:\NetBeansWorkspace\Carica\testInput\books

Work identifier:

Publishing date: 1 1 2013

Images	Name preview	Schemas
Aparicao.jpg		
BandOfBrothers.jpg		
CincoLagoNegro.jpg		
DaVinciCode.jpg		
DDay.jpg		
Equador.jpg		
FellowshipOfTheRing.jpg		
HarryPotter.jpg		
Hobbit.jpg		
LeninStalinHitler.jpg		
Millenium1.jpg		
MinasSalomao.jpg		
Nemesis.jpg		
PegasusBridge.jpg		
Peopleware.jpg		
PlotAgainstAmerica.jpg		
Retribution.jpg		
WarAndPeace.jpg		

☐ Specify the destination folder (If unchecked, images will be stored in the origin folder)

Destination direc... C:\Users\João Cardoso

Figure 4: Carica main dialog

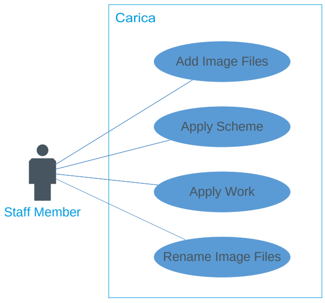


Figure 5: Carica's use cases

3.1.1 *Use Cases in Carica*

There are four use cases that define interactions between a user and Carica. The four use cases are depicted in figure 5.

- **Add image files** Users may add image files to Carica. Once added the user may remove them or rename them, which implies having applied a schema or a work;
- **Apply schema** Users may apply a schema to a batch of image files. This implies having created a schema or edited an existing schema;
- **Applying work** Users may apply a work to a batch of image files. This implies having created a work or edited an existing work;
- **Rename image files** Users can rename batches of image files, by applying schemas or works;

3.1.2 *Carica Application Architecture*

The Carica software application was developed using the Java programming language[8], using a multi-tier architecture, which in this particular case holds three tiers, as can be seen in figure 6.

The data access tier holds the classes that implement domain classes for the application, such as schemas, names, batches and works. The service tier classes on the other hand interacts with the data access tier, by using its classes to implement the logic behind all the functionalities Carica provides. Finally the presentation tier classes implement a graphical user interface with which the user can interact and take full advantage of Carica's functionalities.

3.2 Metadata Editing

The P1.3 presented three primary issues, which were:

[8] Oracle website for Java developers: `http://www.oracle.com/technetwork/java/index.html` Retrieved 2013-10-08

Concept	Description
schema	A schema is a set of options, which derive from BND's naming convention, and that are later applied to a batch of image files, thus renaming them
Work	A work is a set of schemas that are applied in an aggregated fashion to a batch of image files. Works are best suited for processing batches of image files derived from publications, because the structure of the publication remains unaltered regardless of the volume

Table 2: Carica's main concepts

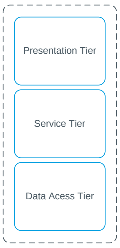


Figure 6: Carica multi-tier architecture

- Two different workflows existed for creating PDF files and enriching them with descriptive metadata;
- The version of the software application ContentE was outdated;
- The staff at the digitizing and image service were not specifically trained to work with ContentE.

The first issue resulted from the fact that only a single Adobe Acrobat licence was held by the digitization and image service at Hemeroteca. This implied that only a staff member was allowed to perform both those tasks using the Adobe Acrobat software tool, whilst the remainder of the staff members had to resort to freeware software applications, respectively JPEGToPDF and BecyMetaPDFEdit. The second issue on the other hand was a consequence of the lack of dedicated IT support to the digitization and image service. This meant that the software application ContentE had never been updated from the originally installed version 1.6, and was therefore outdated. The third issue was the lack of training on the use of ContentE, which implied that the staff at the digitizing and image service could not take full advantage of the functionalities provided by ContentE.

The current version of ContentE (v3.6) is able to both generate PDF files and enrich them with descriptive metadata. This meant that a solution to the first two issues could be achieved if the ContentE version in use at the digitizing and image service was to be updated to the current version. This task was done and the first two issues were solved in this fashion. Additionally Gilberto Pedrosa volunteered to provide assistance in both installing the new version and training staff members in its use, thus solving the third issue.

The optimized P1.3 (see figure 7) is therefore completely different from the original P1.3 presented in section 2.2.2. The whole sub-process is performed through the use of ContentE, and implies three tasks. The first task is the creation of PDF files from the existing JPEG files, so that each PDF will reflect the structure of the originally digitized publication. The second task is the editing of the descriptive metadata on each of those PDF files. The final task is to create a structural metadata record which describes the structure of the JPEG image files in order to replicate the structure of the original publication. The outcome of this operation is the creation of a publication copy in XHTML format, which will later be published in the P1.4.

3.3 Publishing

In what refers to the P1.4 there were three issues to solve, which were mentioned in section 2.3. The issues were:

- The sub-process relies heavily on manual procedures;
- There are no automated tasks;
- Manual editing of HTML files leads to syntactic errors and overall lack of consistency.

The challenge was to solve each of three issues by developing a solution which would have to consider the limitations imposed by the digitizing and image service context. This would imply implementing a system which had to be both easy to use and maintain by the staff members of the digitizing and image service.

Obviously the ideal system for this sub-process would be to implement a relational database in which the records that make up the authority file for the digitizing and image service would be stored. Then a software application equipped with a graphical user interface would be used to create the necessary indexes in the form of HTML files, from the records stored within the relational database. Unfortunately since the staff members at the digitizing and image service have only enough knowledge to be able to interact with computerized systems as users, this ideal solution would not work given the current context.

A possible solution to this problem was the implementation of a Persistent Uniform Resource Locator (PURL) system similar to the already existing PURL.pt[9], which is currently in use at the BND. This sort of system would be able not only to manage the publishing of new works, but also generate the necessary indexes that feature at Hemeroteca's website.

This solution was initially explored but later abandoned for it was found that the bibliographic records were not attached to the digitized works. That implied the development of a complex application to retrieve bibliographic records to be attached to the existing PURL system. This fact proved to be the tipping point between what would have been a viable solution and what would prove to be too much effort for a marginal gain. Additionally the ever pending constraint of the staff members at the digitizing and image service not being able to manage complex computerized systems, meant that this solution as intelligent as it was would not be functional.

[9] PURL.pt webpage (2013-01-03): `http://purl.pt/index/geral/PT/index.html` Retrieved 2013-01-03

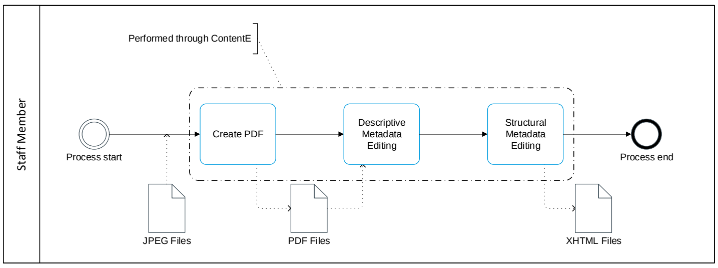

Staff Member
Performed through ContentE
Process start
Create PDF
Descriptive Metadata Editing
Structural Metadata Editing
Process end
JPEG Files
PDF Files
XHTML Files

Figure 7: The optimized Sub-Process Metadata Editing (P1.3)

Therefore a compromise had to be made in terms of the adopted solution. Records would be stored not on a relational database but on Microsoft Exel XLS file which would serve as a *de facto* authority file. This of course would bring about some concurrency issues. In case two distinct staff members happen to edit the XLS file at the same time, conflicts might arise. These issues had to be contemplated in the optimized P1.4.

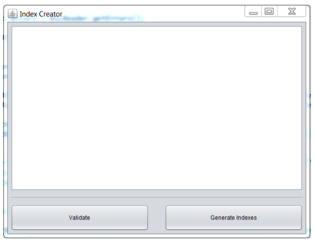


Figure 8: Index creator application

The use of an XLS file as a *de facto* authority file would be the starting point for automating the index creation process. For a centralized repository holding all the information, meant that indexes could be easily generated. All the staff members needed to do was to create a XLS collaboration file for each newly digitized work, indicating which authors collaborated in that particular work. The index generation was to be automatically performed through an index creator application developed using the Java programming language (see figure 8). This application would parse all the existing XLS collaboration files and update all the necessary entries at the XLS authority file, as well as generating all the necessary HTML files that made up the indexes. The idea was to keep staff members at the digitizing and image service from having to manually edit HTML files. This solution would not only solve all three of the issues mentioned in section 2.3, but also add value to the sub-process.

3.3.1 *Optimized Sub-Process Publishing (P1.4)*

Considering the adopted solution, the optimized P1.4 would be composed of three tasks, as can be seen in figure 9.

The collaboration file creation task is a sub-process on its own (see figure 10). It starts by asking the staff member to verify whether the work to be added already exists at the XLS authority file. If it does not, then the first task is to place the work's thumbnail in the correct folder, followed by the creation of a new work entry at the XLS authority file. Regardless, the next task is to compile a list of authors which collaborated in the work that is being processed. The next step is for the staff member to verify whether all the listed authors are present at the XLS authority file, adding them if not present. The last task is the creation a new XLS collaboration file containing the work identifier, and identifiers for all the collaborating authors.

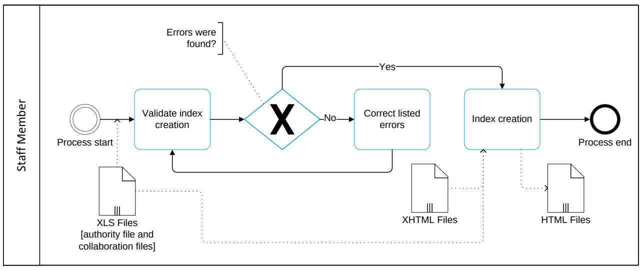


Figure 11: The Sub-Process Index Creation (P1.4.2)

The second task refers to the actual Sub-Process Index Creation (P1.4.2) (see figure 11). Its first task is to validate the index creation, this means running the index creation application and performing a check run looking for possible errors in both the XLS authority file, and all the XLS collaboration files. If errors are found, then they must be corrected and the validation task must be performed again. Once the validation shows no errors, the index creation task is executed and all the HTML files are created.

The last task of the optimized P1.4 to be performed is sending all the files which were created to the remote server which is managed by the DMSI-DASI. A task which is accomplished by using a safe connection through the SFTP, SCP and FTP client for Microsoft Windows WinSCP (see section 2.1).

3.3.2 *Software Applications*

There were two software applications that had to be developed for the optimization of the P1.4.

The first software application was simply a data recoverer which was meant to recover all data stored at the currently existing indexes, and generate new ones based on the sub-process tasks previously described. The development of this application was a lengthy process due to the lack of format consistency of the existing HTML index pages.

The second software application to be developed was intended to be instated permanently within the publishing sub-process as an index creator application. It has two use cases, as can be seen in figure 12.

- **Validate index creation** Users may perform a validation on the artefacts needed to create indexes (XLS authority file and XLS collaboration files), in order to

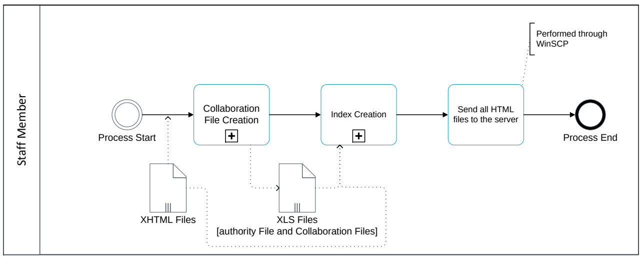

Staff Member
Performed through WinSCP
Process Start
Collaboration File Creation
Index Creation
Send all HTML files to the server
Process End
XHTML Files
XLS Files
[authority File and Collaboration Files]

Figure 9: The optimized Sub-Process Publishing (P1.4)

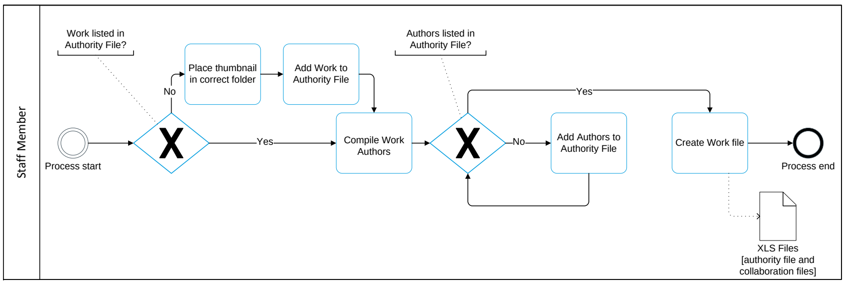

Staff Member
Work listed in Authority File?
Place thumbnail in correct folder
Add Work to Authority File
Authors listed in Authority File?
No
Yes
Process start
Compile Work Authors
No
Add Authors to Authority File
Yes
Create Work file
Process end
XLS Files
[authority file and collaboration files]

Figure 10: The Sub-Process Collaboration File Creation (P1.4.1)

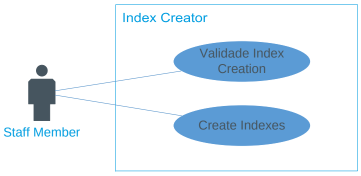


Figure 12: Index creator application use cases

check if there are no errors that would impede the indexes from being created;

- **Create indexes** Users may create indexes. An operation which generates all the necessary HTML index files.

The index creator software application was developed using the Java programming language, using a multi-tier architecture, just as was done in the case of the Carica software application. The application is structured in three tiers, which can be seen in figure 6.

The data access tier holds the classes that implement domain classes for the application, such as authors and works. The service tier classes on the other hand interact with the data access tier, by using its classes to implement the logic behind the two functionalities the index creator application provides. Finally the presentation tier classes implement a graphical user interface with which the user can interact to both validate index creation and to create indexes.

4. CONCLUSIONS

It is always tough to draw conclusions on a thesis project whose execution extended for close to nine months. One could say that the overall objective of the thesis project was achieved, for solutions were found and implemented that resulted in optimizations to a digital library's business processes and consequently to its workflows.

In what refers to the particular case of Hemeroteca's digitizing and image service business process, the issues pointed out in section 2.3 were all met throughout the execution of the thesis project. Their solutions were sometimes not ideal but the deviations from the ideal solution were always brought about by constraints from having to solve a real problem. Nonetheless the staff members of the digitizing and image service can now go about their work on more effective and efficient manner.

The optimization of the P1.2 occurred without any significant problems. This was a bit unexpected, because although the overall objective of the Carica software application was quite simple, the technical challenges to achieve it were rarely so. Obviously this process while mostly uneventful was far from smooth. There were delays during its execution, most of which were down to the technical inexperience of the student to whom this master thesis refers.

Another point worth mentioning was the optimization of the P1.4. The implementation of the solution designed for this sub-process was a lengthy affair. As mentioned in section 3.3.2 the solution implied implementing two software applications. One to recover the existing data, and a second one which would be integrated into the P1.4.

Unfortunately data recovery soon turned into a quagmire, for the lack of consistency between HTML files slowed the process to a crawl. This is why a task which was supposedly simple, ended up extending itself for close to three months. This delay on the conclusion of the execution phase of the thesis project meant that the testing phase of the index creator application could not be as extensive as would have been ideal. Which means that support will have to be given to the staff of the digitizing and image service long before the term of this thesis project has ended.

On balance one can say that the quality of the work developed within the realm of this thesis project was satisfactory. However given the circumstances, and considering that all objectives were achieved, one can not in good faith criticise the commitment to the tasks at hand. This is not to say that given more time both better solutions and an overall greater quality of work couldn't have been achieved.

5. ACKNOWLEDGMENTS

During the development of this thesis I was mightily helped by both my supervisor Prof. José Borbinha, and my *de facto* co supervisor Gilberto Pedrosa. It was a pleasure working with you both, thank you.

6. REFERENCES

[1] Borbinha, José and Gil, João and Pedrosa, Gilberto and Penas, João". The Case of the Digitized Works at a National Digital Library. In *Proceedings of the 2nd Int. Conf. on Document Image Analysis for Libraries*, DIAL '06, pages 116–125, Washington, DC, USA, 2006. IEEE Computer Society.

[2] Borbinha, José and Pedrosa, Gilberto and Penas, João. ContentE: flexible publication of digitised works with METS. In *Proceedings of the 9th European Conf. on Res. and Adv. Tech. for Digital Libraries*, ECDL'05, pages 537–538, Berlin, Heidelberg, 2005. Springer-Verlag.

[3] José Borbinha, João Gil. Regras para Estruturas de Directórios e Nomes de Ficheiros de Imagens Digitalizadas.

[4] Network Working Group. RFC 1512: MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies. `http://tools.ietf.org/html/rfc1521`, 1993.