DIGITIZING EGYPTIAN NATIONAL DOCUMENTS ARCHIVE: CHALLENGES AND SOLUTIONS

Ahmed Samir[1], Bassem Elsayed[1], Noha Adly[1,2], and Magdy Nagi [1,2]
[1]Bibliotheca Alexandrina, El Shatby 21526, Alexandria, Egypt
{ahmed.samir, bassem.elsayed, noha.adly, magdy.nagi}@bibalex.org
[2]Computer and Systems Engineering Department,
Alexandria University, Alexandria, Egypt

1. Abstract

The Egyptian National Documents Archive (Dar El-Mahfouzat) dates back to 1805, to Mohamed Ali's era. This renders the archive a crucial and valuable aspect of Egyptian heritage. Dar El-Mahfouzat has partnered with the Bibliotheca Alexandrina (BA) for creating a digital archive for a collection of documents comprising more than six million pages physically pertaining to the Archive for digital preservation and access. In implementing this project, the BA has undertaken three main aspects simultaneously; building the digitization facility, developing and installing the necessary software tools and building the capacities of the staff. Each aspect represented a challenge accompanied with the notion of deploying state-of-the-art technologies for a proficient digital output. Two million pages have been digitized to date and the digital archive has been created to support multilevel authorization to secure access to classified documents.

This paper will elaborately reflect the work done by the BA in creating the digitization workflow, and developing the associated tools for archiving digitally a grand historical facility in Egypt, while demonstrating the challenges encountered and their handling.

2. Introduction

Having to deal with a huge number of material, the team started with analysing the selected collection of documents to be digitised, thus assessing the size of the material to handle and accordingly the needed equipment to be installed. Therefore, an efficient workflow system is required to organize the operations involved. And this includes, inter alia, coming up with a flexible metadata schema for metadata extraction that would render the digitization workflow expandable for any types of documents. The core is the digitization workflow used in this project has been derived from the BA's Digital Assets Factory (DAF). DAF is an open source system built in-house in which the Library's digital assets are created and maintained. In the framework of this project, DAF has been customized and integrated into the digitization workflow of Dar El-Mahfouzat documents.

Yielding a digitised output, the workflow includes a metadata/digitized output matching scheme where digital files are combined with their metadata for conformity and further categorization. In order to implement this workflow efficiently, it has been essential to build the human capacity to professionally operate the workflow. This represented a challenge in terms of implementing an efficient applicable methodology to ensure a premium quality output.

The paper presents a detailed review of the implementation of Dar El-Mahfouzat project, and discussing its main components and the implemented workflow, giving a detailed overview on the three main processes of the system; the metadata process, the digitization process and the final matching process. As the cornerstone of the digitization workflow, the Digital Assets Factory is thoroughly explained giving a full account on its basic components and modules, and showing how it has been reconfigured to act as the core of the digitization workflow at Dar El-Mahfouzat project. Finally, an overview is given on the components of the on-site constructed digital laboratory and building the human capacity.

3. Objectives

The main objectives of Dar El-Mahfouzat Digital Archive are:

- Preservation. The preservation of documents by reducing wear and tear on the originals for reference and reproduction.
- Increased accessibility. Facilitating their future exploitation by a broader number of researchers and interested parties
- Added value. Enhancing users' understanding of the documents archival context, identifying the documents keywords and linking the documents with their related events, places and persons.

4. Defining the main components of the project

4.1 Setting the physical infrastructure

Building a facility to accommodate the digitization equipment represented the first challenge taken on by the BA. Servers, scanners, storage nodes, laboratory machines, electric power cabling and building the networking infrastructure, were studied, planned and implemented. It will be discussed in details later in this paper.

4.2 Data Analysis

The archive comprises millions of official Egyptian documents varying from birth and death records, employees files, land ownership certificates, governmental correspondences and other collection types. The documents collection contains Arabic and English documents, contains both hand written and typed documents, contains both colour documents and black and white documents. Good metadata makes it possible to catalogue and effectively present and retrieve digital information. As the metadata is a fundamental element of any digitization project, the *Metadata extraction* is the main process of the data analysis for the Dar El-Mahfouzat documents collection. The collection is divided into two main categories. The first is multi-document files (such as employees files), and these are called DOC category. The second category is single-document type (such as birth and death certificates and these are called REGISTRY RECORDS category. The challenge in this process is how to deal with the large number and varieties of metadata fields for the different types of documents. This led to the need for designing a flexible schema. Defining and selecting all the metadata fields available for each document type was not easy due to the diversity of documents . All the shared metadata fields were normalized into the main document record and a list of different metadata fields was defined and linked to each document type, thus appearing in the metadata entry applications upon selecting a specific document type. These metadata fields typically describe the document content, format, source and many other attributes.

4.3 Software Development

The BA developed the applications needed for the workflow processes which divide into two groups:

- The metadata applications which handle the metadata entry, review and final matching processes.
- The digitizing and archiving applications which will handle the digitization process.

These applications will be discussed in details in the implementation section. Using the Digital Assets Factory (DAF) in the digitization process required developing crucial customizations which are elaborated in DAF customization section.

Authorization and Information Security: For any archive, the access rights issue is very critical. The issue of access levels for individuals handling these documents, emerges the need for supporting a multilevel authorization for the working staff to secure the classified documents. This is implemented in the workflow using an access matrix for users privileges and assigning the needed privileges only for each user. All users actions are recorded and audited throughout the workflow.

5. Implementation

5.1 Workflow Overview

Following digital archiving best practices and metadata standards, the digitization workflow (as shown in figure 1) consists of 3 main processes:

- Metadata process where the physical documents are labelled, described and linked together using keywords
- DAF (Digital Asset Factory) process where the documents are digitized in a predefined workflow according to their types
- Matching process where digital files are coupled with their metadata for conformation and further categorization.

The following section gives a detailed overview on each of these processes and discusses the integration between them, which has been quite a challenge.

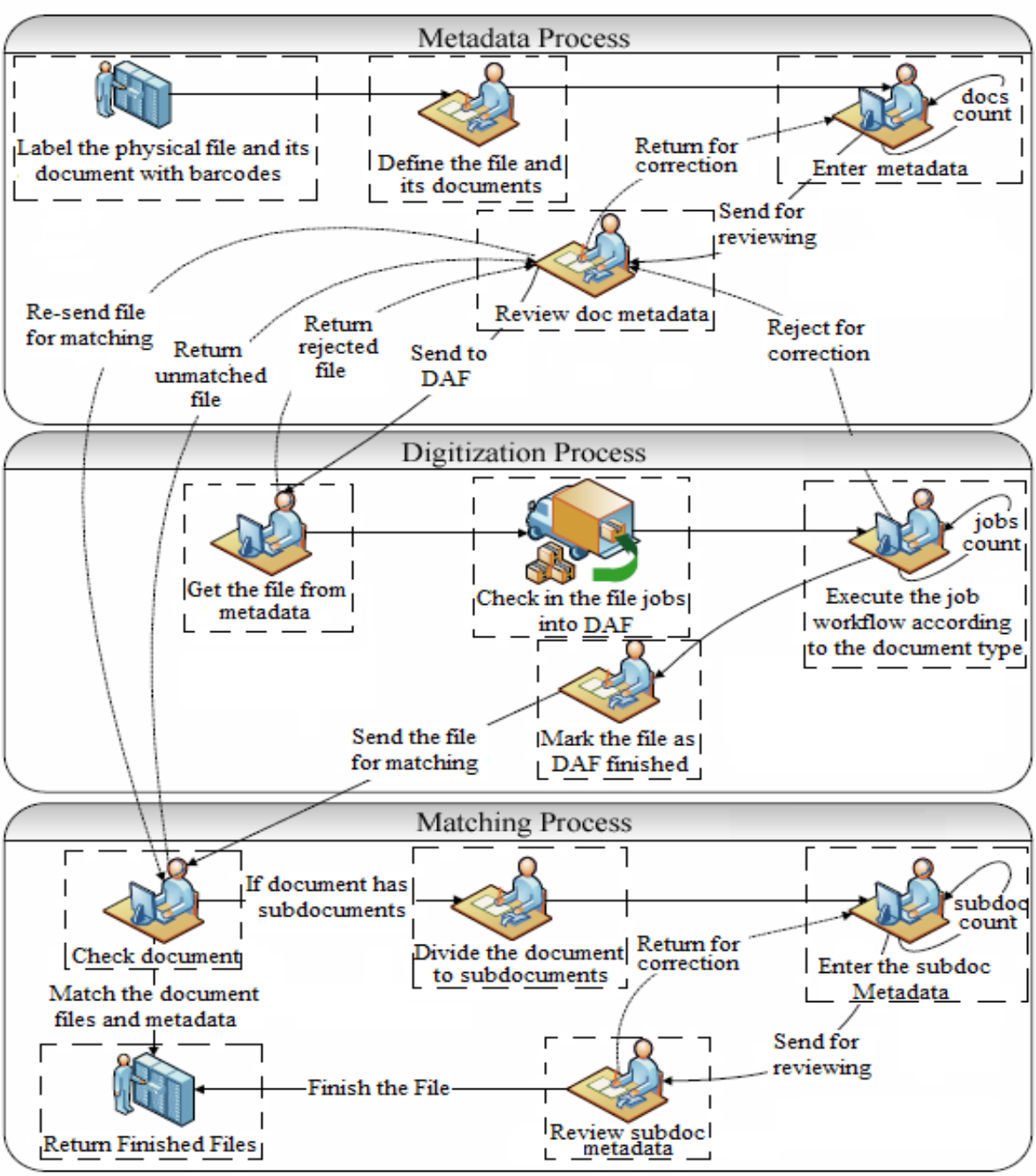

Figure 1. Digitization Workflow

In case of multi-document files (DOC category), each document is identified as a unique digitization job. The file is divided into subdocuments (which represent a group of pages inside a document), where each has its own metadata fields that describe and identify it.

5.2 Metadata Process

This process is divided into two main units:

5.2.1 Classification Unit

The classification unit is the document entry point into the workflow. The collection files are classified according to their collection type which would be assigned and associated with them throughout the workflow. The collection handled at Dar El-Mahfouzat has 14 types including; forensic books, public shops registration books, population statistics books, the incidence of slavery books, substitutionary freedom tickets books and the expropriation books. The collection is defined as files and each file consists of one document in case of REGISTRY RECORDS category such as birth and death certificates files or consists of multiple documents in case of DOC category such as seals fingerprints books.

As shown in Figure 2, first, the file as well as each document in it are given a barcode label. This facilitates and accelerates the document identification and tracking throughout the workflow by using a barcode reader. The classification unit generates a serialized barcode based on a pre-defined naming convention.

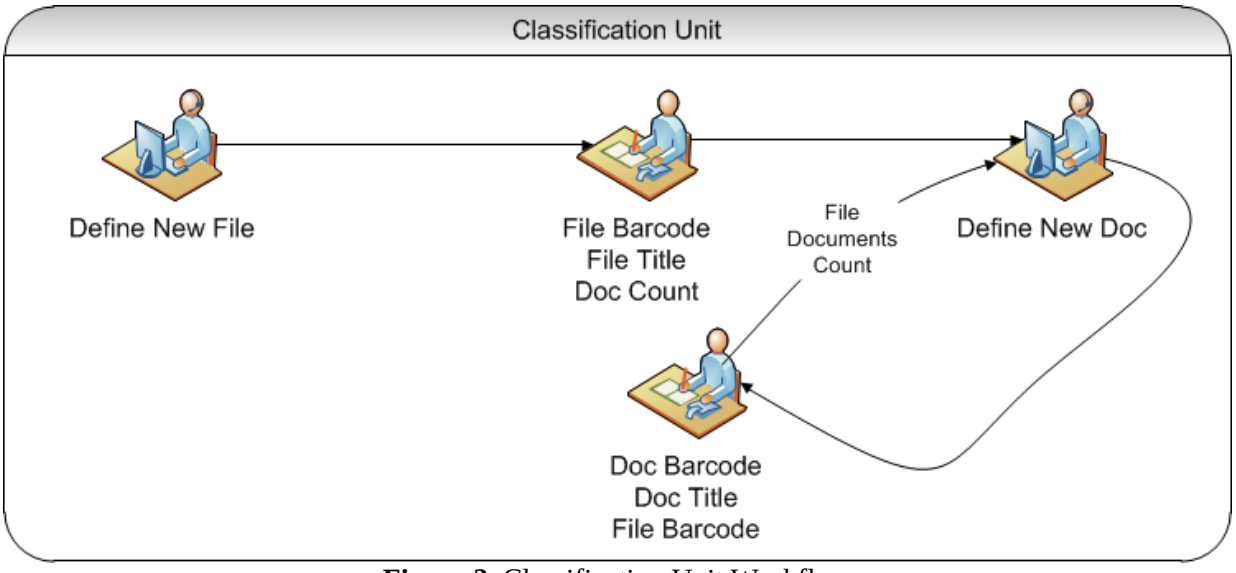

Figure 2. Classification Unit Workflow

5.2.2 Metadata Unit

Metadata unit consists of two processes:

5.2.2.1 The metadata entry process

As shown in figure 3, Each document in the file is described through fulfilling the required metadata fields. A flexible schema has been designed to allow the metadata application display the fields attached to each collection as per its type. For each document a real page

count and an actual page count is entered. The real page count is the supposed page count for this document which was known from the numbering or from the collection type standard format. The actual page count describes the existing document page count. After the document metadata is entered, it is then marked ready for reviewing.

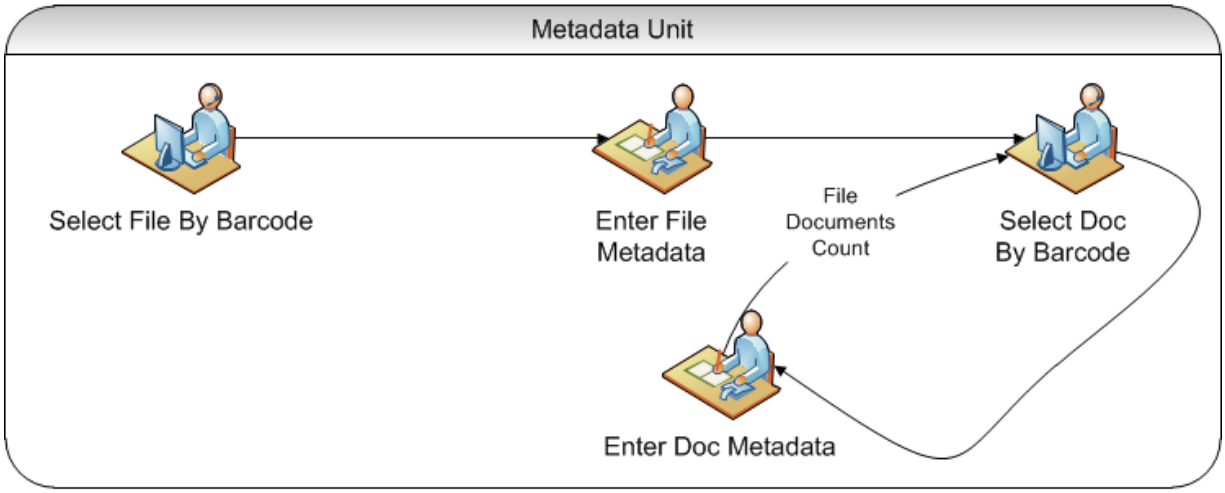

Figure 3. Metadata Unit Workflow

5.2.2.2 The reviewing process

This is the phase where the reviewer checks and revises the entered metadata. The reviewer has the ability to modify the metadata when necessary. Then, he/she takes a decision regarding whether to return the document to the metadata entry process to be corrected or mark the document as "Finished". After all the documents in the file are "Finished", the parent file is automatically marked as "Finalized" as well, and it becomes ready to move to the digitization process.

5.3 Digitization Process

5.3.1 DAF Overview

The Digital Assets Factory (DAF) is a software developed by the BA to be used for automating and controlling the digitization process. DAF provides a configurable and flexible management tool for any digitization workflow where several workflows can be configured for different types of digital objects. When digital objects to be ingested have their metadata in an external repository or a database, DAF integrates with these external sources of metadata through the development of plug-ins.

Digitizing different types of objects requires the incorporation of several tools, such as software tools for scanning, image processing and OCR .There is no one-size fits-all tool that can perform all the different digitization tasks, where each object type might rely on very different set of tools and a workflow process for digitization. DAF integrates with these tools to assist repository administrators in managing the workflow.

The system supports different workflows for different types of material including the scanning of the textual material, the processing of these scanned files to enhance their quality, performing the Optical Character Recognition (OCR) on textual material and encoding the digitized output by generating a version suitable for publishing. DAF provides a database system to keep track of the digitization process. It also keeps track of the materials to be digitized and provides timely reports on various levels of management describing the workflow on daily, weekly or longer basis and allows online queries about the current status of a certain document in the digitization process.

As shown in Figure 4, DAF divides each digitization workflow into phases. It can integrate with automated tools and scripts, checking their status at each phase and verifying their output through pre-

phase and post-phase checks. DAF also makes sure the output is compliant to the digitization standards in terms of file types, number of files and naming conventions thus minimizing human operators' tasks to only what humans are good at: OCR correction. A Reporting Module provides timely reports about the status of the digitization operators and the automated tools. BA provides DAF to the community as an open source tool (http://wiki.bibalex.org/DAFWiki)

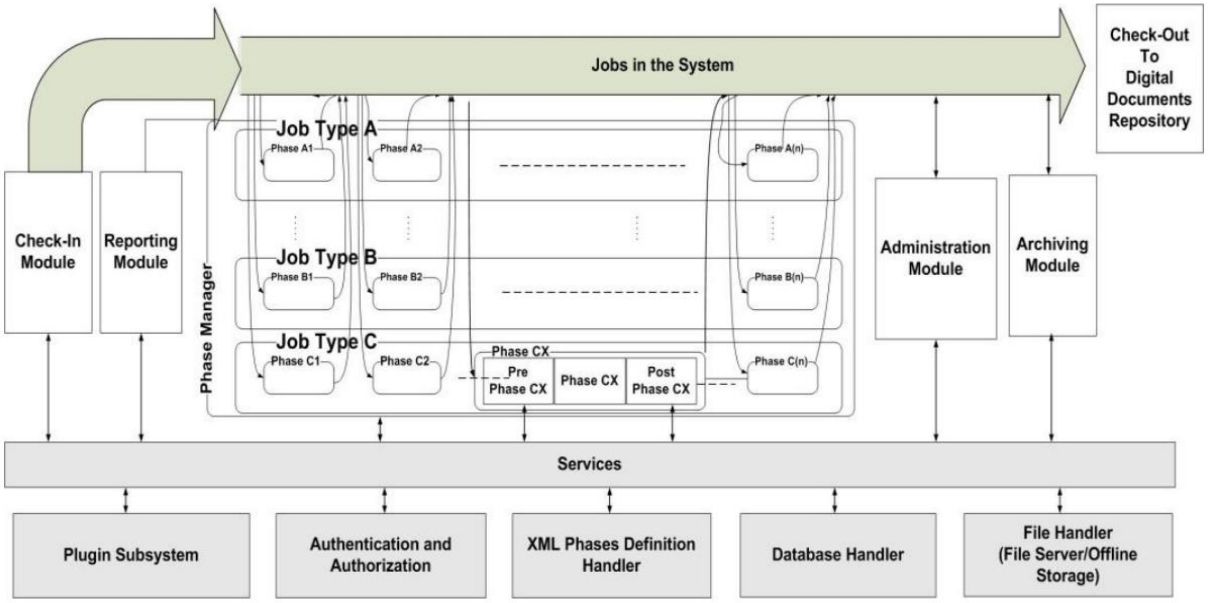

Figure 4. DAF architecture

5.3.2 Definitions

5.3.2.1 DAF Job

It is the main entity that represents the object being digitized. It can be a REGISTRY RECORDS category or a DOC category as discussed in the next point. The phases by which a "Job" passes depends on its type. It has a priority and a life time in the workflow otherwise it will be reported as a Late Job.

5.3.2.2 DAF Job Types

DAF Job Type is the logical grouping of Jobs. For Dar El-Mahfouzat collection, two main job type categories are defined:

- **REGISTRY RECORDS** Category: is the set of files which consist of many pages grouped within a single cover such as Birth and Death records, Forensic reports and land ownership records. A file consists of one document only and will be represented by a single job. This would affect its digitization path and the file would be scanned on the A1 and A2 scanners.
- **DOC** Category: is the set of files which contain several standalone documents such as employees service files which consist of many documents like employee birth certificate, qualification and contract. A file consists of multiple documents and each document represents a separate job.

Assigning the suitable category for each object defines its digitization path throughout the workflow and the scanners used. REGISTRY RECORDS category objects are scanned on A1 and A2 scanners, while DOC category objects are scanned on A3 scanners.

If a document (category) is typed, then it is branched out to OCR job type, while if it is handwritten, then it is branched out to No-OCR job type. As shown in Figure 5, Each job type consists of several phases designed to accommodate the needs of every document type. All the documents in the file are ingested into DAF system as separate jobs each of a certain job type according to the document collection type and attributes.

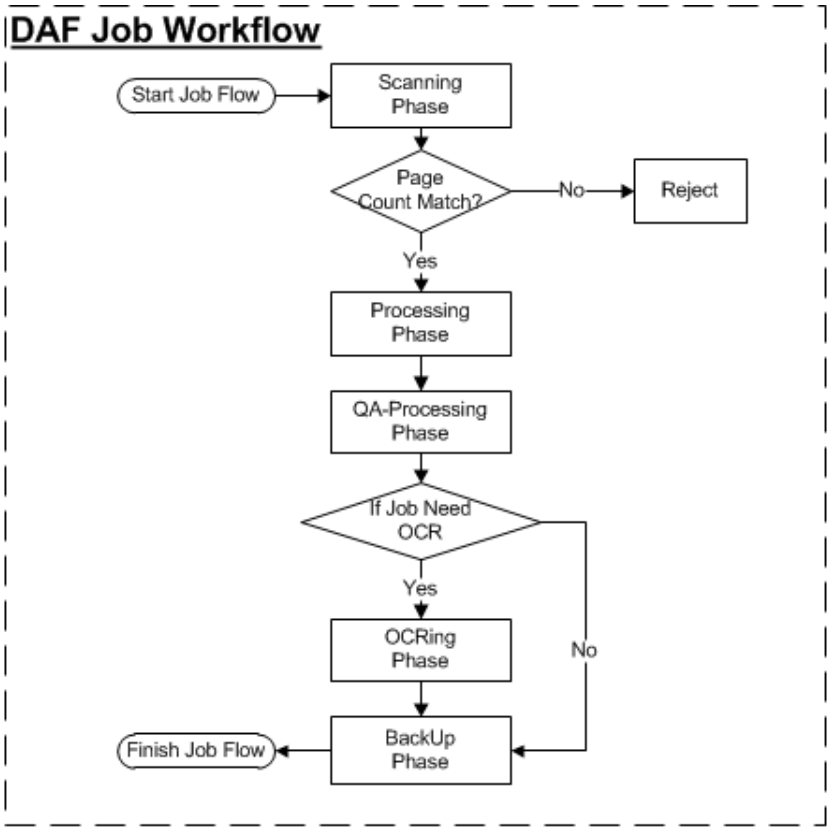

Figure 5. DAF Job Workflow

5.3.2.3 DAF Phase

A Phase represents a task or a unit of work that should be applied on a specific Job Type in its digitization workflow. Each Job Type has its own sequence of Phases defined in a Phase Sequence to obtain a digitized version of the Job. Phases have significant role in applying checks and actions. all Phases have an XML definition that holds checks and actions format.

5.3.3 DAF Main Modules

5.3.3.1 XML Phases Handler

The XML Phases Definition Handler is responsible for interpreting and applying the XML definition of the Phases. Each Phase has its own XML Phases Definition, specifying the prerequisites and actions that need to be done before and after each Phase. The XML definition contains two main sections; Pre-Phase and Post-Phase. Each of these sections is composed of three subsections; Physical, Database and Reflection Call.

Physical section: In the Pre-Phase section, the Physical section allows to describe the necessary folders and files structure required to start work in the Phase and which of them should be copied to the client's working folder to execute the Phase. In the Post-Phase section, the *Physical section* allows to describe the necessary files and folders structure required to complete the Phase. It also defines which of the folders and files should return to the file server.

Database section: It is usually used in the Post-Phase section. It allows to define the structure of database information that should be submitted after finishing the Phase. It contains listing and naming for the fields that should be filled by the operator during his work in the Phase. These fields are saved as XML text, with the Phase information in the transaction log for reference and query later using XPath.

Reflection Call section: This section allows to specify the Java function that should be executed either in the Pre-Phase or Post-Phase. The function can start any process including files management, data entry, zipping, or encoding the files.

5.3.3.2 Authentication and Authorization Handler

All the provided interfaces and services of the system are accessible through an authentication and authorization handler. This handler is responsible for customizing the application interface for the logged in User. Moreover, it authorizes each action or request submitted by User.

5.3.3.3 File Handler

The File Handler component is used by the XML Definition Handler to manage the file copying and movement. It is responsible for the necessary ftp handling with the file server and local file handling on the clients.

5.3.3.4 Database Handler

All the database interactions are done through the Database Handler, which is responsible for interfacing with the database stored procedures. All DAF logic is implemented and saved into stored procedures and functions, there is not any single query passed to the database without calling Stored Procedure. This technique is the key behind DAF flexibility and its adoption imparts security and code flexibility If there is a logical change needed the Java code will not be modified, instead, only the effected stored procedures or functions are modified. Definitely, this scenario is valid for code's segments that need nothing but database changes.

5.3.3.5 Check-In Module

The Check-In Module is responsible for creating a job in the system and fires it to start. Although the Check-in Module determines the first Phase of a job depending on its Phase Sequence , the system allows to handle an exception and allows the Job to start from an intermediate Phase within the workflow as long as its prerequisites are met.

5.3.3.6 Check-In Plug-In Subsystem

DAF check-in has been designed and built to allow for the integration with any metadata source as the Integrated Library System (ILS), document registry, MARC or MODS files. This flexible integration has been achieved by making the check-in module built as plug-in based. The system allows the DAF implementers to write their own check-in plug-ins for their different metadata sources.

5.3.3.7 Administration Module

The Administration Module is responsible for the necessary system parameterization and settings. It allows the administrator to define and manage the Job Types as per their digitization

workflow and phases, the roles of users, workstations, and collections. It also provides the facility to control the matrix covering the relation between users, workstations, job types, and collections. A "Role" is a group of permissions and each user is assigned a single Role. A "Collection of Jobs" is a group of physical Jobs. A "Collection" may contain several documents of different Job Types. The "Workstation" represents the computer where the execution of the phases is performed.

5.3.3.8 Reporting Module

The Reporting Module provides the necessary reports to allow for managing the Jobs within the workflow. The module provides four types of reports:

- *Workflow Tracking reports:* provide the status of the Jobs in the system. It provides for each Job Type, the number of Jobs pending , started, and finished within each Phase. It also can separate the Jobs revisiting a Phase to provide information about the new and old Jobs in the system
- *Pending Items report:* provides the redirected or rejected Jobs by the operators. The administrator can accept or reject the redirection through this report.
- *Late Jobs report:* provides a list of the Jobs that exceeded its due time either within a Phase or in the whole workflow, since there is a due time for each Job and a due time for each Phase.
- *Operators Rate report:* helps the supervisors and higher level management to get the laboratory overall production and provide a tool for evaluating the operators.

In addition to the above reports, the module provides a query builder, where the User can design his/her own report on the Jobs. The module also provides search capabilities for Jobs by the various attributes of the job ID, title, creator, Job Type, and language. This search can be considered as an access point to the Job to assign or retrieve it from an archive.

5.3.3.9 Archiving and Retrieval Module

DAF uses active storage servers during ordinary workflow to upload and download jobs from and to a User's workstation. Ordinary workflow will eventually fill up all space in these storage servers. Archiving data is a solution for this problem; it also offers the opportunity to save data into external media where data may be retrieved from this media in the future. Technically, archiving process input represents the Backup Phase output. In other words, Backup Phase is responsible of zipping and versioning the Job folder (the Job can get more than one version) . Logically, archiving solves the space problem by giving the operator the ability to Check Out Jobs, which means removing the physical files of the checked out Jobs from Storage Server. DAF prevents checking a Job out unless it has been archived.

The Archiving Module displays the Job Versions that have not been archived before, these versions are transferred to the Archiving Storage Servers. Archiving Manager has 2 modules:

- The archiving process core which decides which Jobs Versions need to be archived and on which archiving storage node. It creates a folder named with the job ID and version number containing 3 items:
    - Compressed job version folder
    - File contains the MD5 checksum of the compressed file for verification.
      XML file contains all the job history in the form of transactions log which keeps track of the job in case of its retrieval Then, marks this version as archived
- Check out module: Job lifecycle ends with archiving all its versions on archiving storage. this means that the Job is in a finished Backup phase state. Actions associated with Checking out a Job are:

1. Moving Job's logs entries archived transactions log table.
2. Deleting Job's folder from the Working Storage Server

The retrieval module serves for the retrieval of the Jobs from the Archiving system, either for the jobs that have been Checked-Out from the System and need more processing, or for the jobs that are still in the system but it is preferred to work on their archived version rather than the existing one.

5.3.4 DAF Process in Dar El-Mahfouzat digital lab

The DAF process in Dar El-Mahfouzat project is divided into four steps:

- **Step 1:** DAF Ready application is used as the entry point for DAF process where the file is validated referring to the defined documents of the file before it is inserted into the system. To check that all the file documents are ready to be inserted into the digitization process.

- **Step 2:** files are checked into DAF using Dar El-Mahfouzat Check-In plugin which will be discussed later in this paper. Barcoded, all the documents in this file are added as jobs in DAF with the needed metadata to define their job type. The entered metadata comprises the document's type, title, language, page count and printing type which is whether hand written or printed.

- **Step 3:** The document is moved through the workflow according to its job type, after it passes the scanning phase, it goes into the processing phase so the images are processed for better quality by eliminating the distortion and trimming their outlines. Then, the quality of the scanned files and processed files are reviewed by the Quality Assurance phase, where passing that, the processed files are converted to a light version for the matching process.

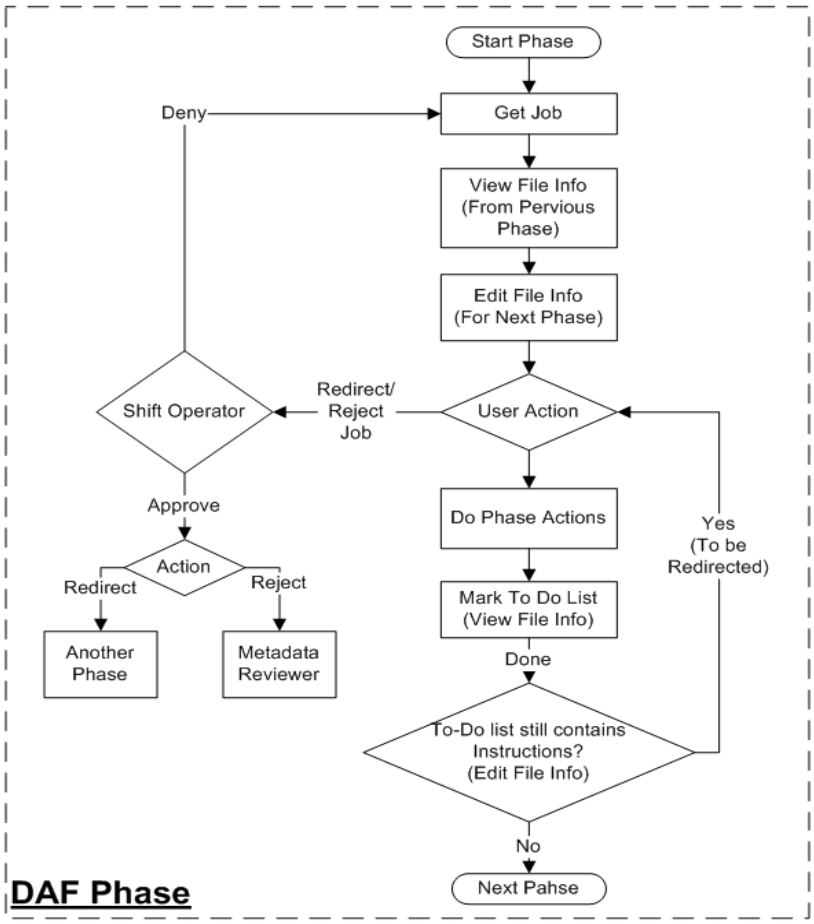

Figure 6. DAF phase workflow

As shown in Figure 6, if an error is reported in any phase, the job is then redirected back to the phase at hand stating the reasons. The shift operator reviews the reason and takes the decision of whether to confirm this redirection or refuse it sending the file back to normal workflow. Thus the digitization flow is flexible for errors correction.

If in the scanning phase the page count of the document does not match the entered count in the metadata or the metadata does not match physical document, then the document is rejected and sent back to the metadata process to be reviewed and handled accordingly. After the document is corrected, the metadata reviewer resends it to DAF entry point where the document is checked-in with its same job ID to continue in the workflow pipeline.

- **Step 4:** Finally the document is archived and marked as "Finished". After all the documents are marked "Finished", the file is also marked "Finished" and it becomes ready for the matching process.

5.3.5 DAF customization

DAF was customized to match Dar El-Mahfouzat digitization workflow, which has impacted a flexible management of the digitization process by adding the following:

5.3.5.1 Dar El-Mahfouzat Check-In plugin

Special Check-In plugin has been developed for Dar El-Mahfouzat Digital Lab collection which reads the required metadata from the metadata database using the barcode labelled on the file. This ingests the file documents into the digitization process as a job for each document. According to the document collection type, it is assigned to the proper job type.

5.3.5.2 Image Conversion Reflection Call

An image conversion reflection call has been developed which is called after the quality assurance phase of the processed document images to generate a light version for the metadata matching process. This light version is generated for faster retrieval and easier review in the final matching process. Also, thumbnails for the processed document images are generated for the further grouping and classification done when creating the subdocuments as described in details in the Matching process subsection.

5.3.5.3 Adding Reject to Metadata workflow path

The BA handled the rejection scenario when the operator finds a mismatching between the physical document and its retrieved metadata, that is when the metadata describes another physical document other than the one in hand or page count mismatch. So, the document is returned to the metadata unit for reviewing and correction then re-ingested into the digitization process. The system traces and keeps a record for each document returned from DAF to the metadata unit, to follow up the progress of the processed file. In order to preserve the physical order of the file documents, the file will not be marked as "Finished" in the digitization process until all its documents are finalized.

5.4 Matching process

The matching process is divided into three steps:

Step 1: The matching step, where the entered metadata fields for the document are matched with the document's digitized images. The document title, source, author, page count and the rest of the defined attributes which are defined for collection type at hand are checked for consistency with the image. If they were not matched, then the document is returned to the metadata process for correction, then sent back to resume the matching process. If metadata fields match the document's digitized images, then, the document moves to the next step.

Step 2: The classification step which is carried out to documents which consist of subdocuments each with unique metadata. In this case, the document is divided into subdocuments using the following steps:

- Creating the subdocument
- Selecting the subdocument images from the document images, marking the thumbnails of the selected images for attaching them with this subdocument
- Entering the subdocument metadata according to the document type where each subdocument has attached metadata fields.

For example, in case of the seals fingerprints books for the Egyptian countryside. Each book may contain a group of adjacent villages. This document must be divided into several subdocuments for defining their regional metadata which would identify and present these group of images for searching in their subdocument metadata fields.

Step 3: The reviewing step, where the subdocuments metadata is reviewed and corrected, then the subdocument is marked "Finished". After all the subdocuments of the file documents are also marked as finalized the file is automatically marked as "Finished" and its light version is archived. At this point, the digitization workflow of the document is complete.

5.5 The complete project workflow

Figure 7 shows the file and its documents life cycle across the workflow.

6. Building the Human Capacity

As digital archives require a professional staff with the knowledge and skills to meet the needs of the digitization workflow applications, building the human capacity for Dar El-Mahfouzat digital archive was quite a challenge. The staff selected for the project were qualified with at least a basic level of IT competency in a Windows environment. A training program was delivered to selected candidates aiming to foster key skills for digitization including image scanning to capture a digital image from a physical object, image processing and applying OCR (Optical Character Recognition) techniques which convert imaged text into machine-readable format.

Additionally, more specific trainings were conducted on the software at hand as per the user's assigned tasks such as: cataloguing and metadata applications; classification application, metadata entry and reviewing application, matching application for the metadata process operators and DAF (Digital Asset Factory) application for the digitization process operators.

Documentation on the workflow tools had been provided to the operating staff as well as the necessary user manuals for the developed applications describing all the feature of each application in details and with sample snapshots.

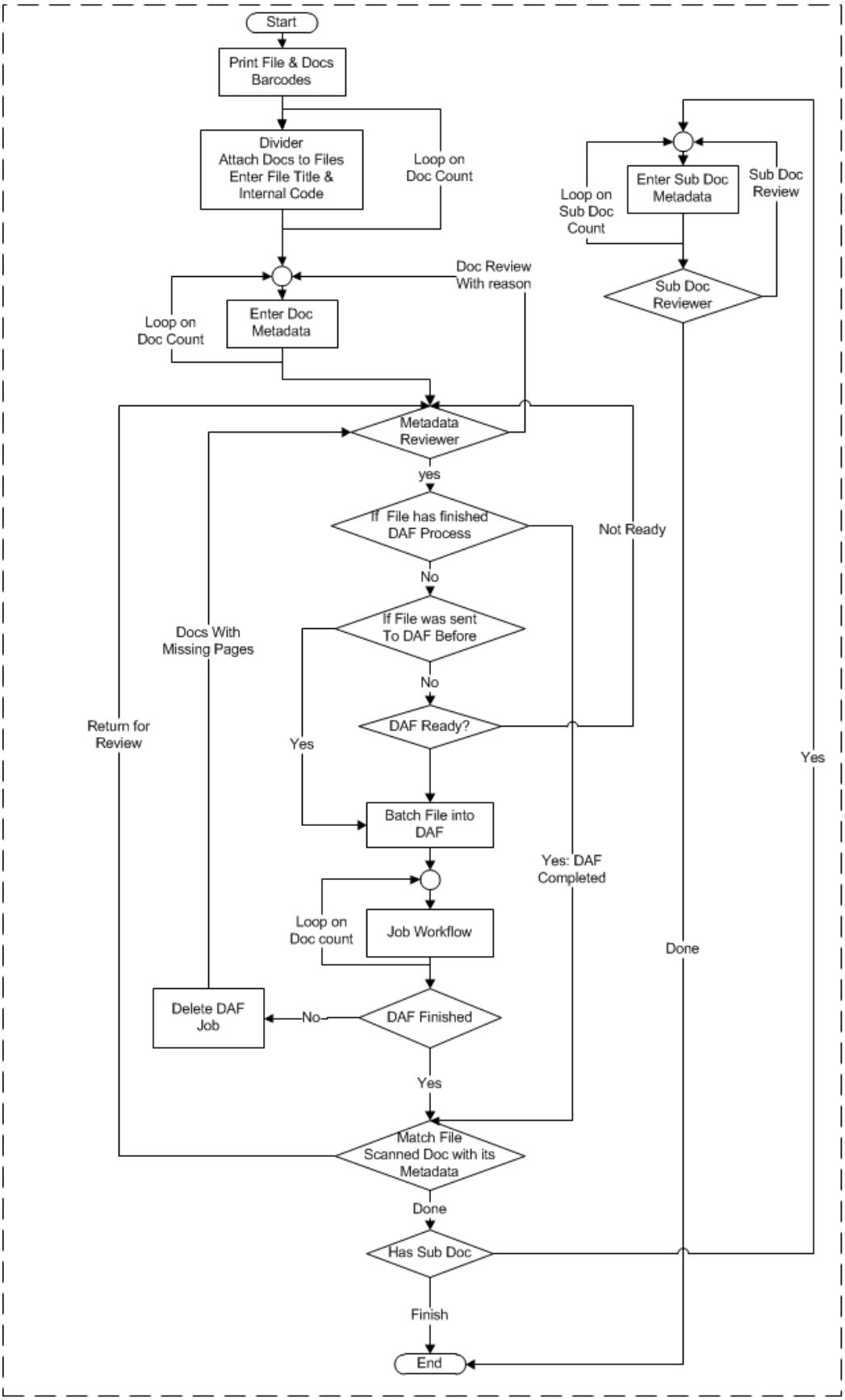

Figure 7. DMF Workflow with the file and document tracking

7. Building the Digital Laboratory

Building a facility to accommodate the digitization equipment represented the first challenge taken on by the BA. The construction of Dar El-Mahfouzat Digital Lab has been divided into two main phases:

- **Phase 1** included assessing the required hardware which would accommodate ingesting this huge collection of documents. Accordingly, four main components were identified:

  - The first component comprises five powerful servers including; an application server for hosting the metadata and DAF databases, a Domain Controller server for users authentication in addition to another one for failover, ISA and Proxy server for secure access to the internet and a file sharing server to be used among users for document sharing and backup.

  - The second component is the storage nodes which were built to accommodate the huge size of the collection which comprises more than six million pages. The node consists of a rack of 40 beta nodes each of 4Tb storage thus yielding a storage capacity of . 160Tb. The overall capacity is equally divided into two sections of 80 TB each; an operational storage and a backup storage The operational storage is further divided into a 20Tb section for active jobs and a 60Tb section for archived ones. we estimated for the six million page as for each document there will be one or more version, each version consists of the scanned and processed images for this document. Then this version is compressed and archived.

  - The third component is the lab machines. Machine requirements were assessed based on the accommodated needed software requirements to yield high quality output. 50 machines were availed and divided to serve the staff working on metadata processing and those handling the digitization process. Each group of staff are working into 2 shifts each of 25 operators.

  - The fourth component is the scanners: regarding to the files different dimensions, we needed to provide the following: an A1 scanner, three A2 scanners and five A3 scanners to handle the variety of files dimensions.

- **Phase 2** included building the network infrastructure on site. It is worth noting that Dar El-Mahfouzat premises is an ancient historical building, thus building the network infrastructure represented quite a challenge given the design and general environment of the building. BA rebuilt the electric power cabling and established the network, installed the hardware, configured the servers and deployed the workflow applications on the lab machines according to the users access matrix and assigned tasks. (see figure 8)

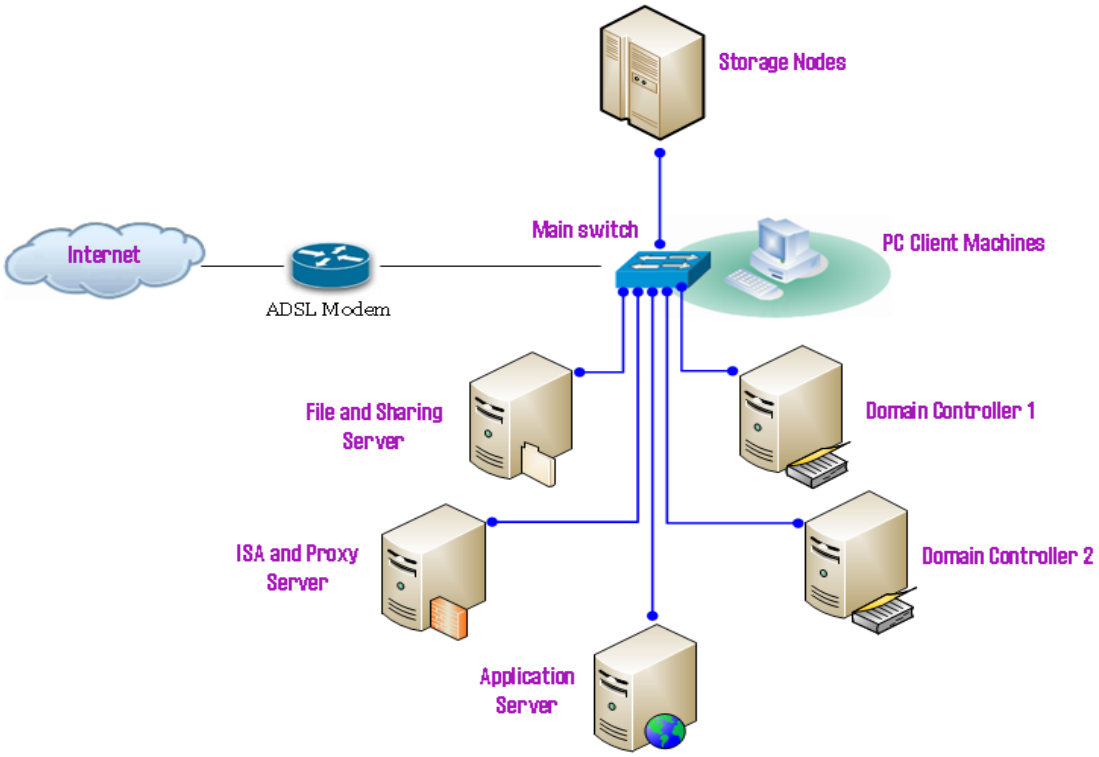

Figure 8. DMF Network Architecture

8. Conclusions

The role of Bibliotheca Alexandrina in the establishment of the Egyptian National Documents Digital Archive (Dar El-Mahfouzat), breaking through all these challenges and applying state-of-the-art technologies and standards, gives a leading model for the digitization projects not only in Egypt but for the developing countries worldwide. Throughout this project, the BA's Digital Assets Factory (DAF) has further proved to be a configurable and flexible management open source tool for any digitization workflow.