DIGITAL PRESERVATION: HANDLING LARGE COLLECTIONS CASE STUDY: DIGITIZING EGYPTIAN PRESS ARCHIVE AT CENTRE FOR ECONOMIC, JUDICIAL, AND SOCIAL STUDY AND DOCUMENTATION (CEDEJ)

Ahmed Samir
Bibliotheca Alexandrina
El Shatby 21526
Alexandria, Egypt
ahmed.samir@bibalex.org

Ahmed Sharkas
Bibliotheca Alexandrina
El Shatby 21526
Alexandria, Egypt
ahmed.sharkas@bibalex.org

Noha Adly
Bibliotheca Alexandrina &
Computer and Systems Engineering Department,
Alexandria University
Alexandria, Egypt
noha.adly@bibalex.org

Magdy Nagi
Bibliotheca Alexandrina &
Computer and Systems Engineering Department,
Alexandria University
Alexandria, Egypt
magdy.nagi@bibalex.org

Abstract

Managing the digitization of large collections is quite a challenge not only in terms of quantity, but also in terms of text and material quality, designing the workflow system which organizes the operations, and handling metadata. This has been the focus of the Bibliotheca Alexandrina during its partnership with the Centre for Economic, Judicial, and Social Study and Documentation (CEDEJ), to digitize more than 800,000 pages of press articles dating back to 1976. This triggered a need to design a workflow to manage such a massive collection proficiently. This required simultaneous intervention of four main aspects; data analysis, developing a digitization workflow, implementing and installing the necessary software tools for metadata entry, and publishing the digital archive.

This paper demonstrates the workflow system implemented to manage this massive press collection, yielding more than 400,000 items to date. It illustrates the BA's Digital Assets Factory (DAF); the nucleus of the digitization process and the tools and stages implemented for ingesting data into the

system. The outflow is also discussed in terms of organizing and grouping multi-part press clips, in addition to reviewing and validating the output. The paper also discusses the challenges of associating the accessible online archive with a powerful search engine supporting multidimensional search.

Keywords: Digital Preservation, Bibliotheca Alexandrina, Workflow

Introduction

Digital preservation is considered one of the most important targets for cultural and historical documentation. This paper tackles the challenge of handling large collection preservation through digitization; the paper will illustrate the main challenges of such process and the need for designing a digitization framework to be applied for all large collection. The digitization framework components are to be explained and discussed. The paper presents a detailed review of the implementation of CEDEJ press collection as a practical case study in which BA has applied the proposed digitization framework. Finally we will discuss the planned future work and enhancement that will be implemented on BA digitization framework.

Objectives

The main objectives of BA Digitization framework are:

- Efficient Handling of large collections, making the digitization process organized and easy to track and manage smoothly
- 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.

Digitization Framework Design

Framework Overview

Following digital archiving best practices and metadata standards, BA designed the processes needed for the digitization framework which are divided into four main processes (as shown in figure 1):

- The indexing process in which the collection are indexed and initially categorized and prepared for injecting into the next process

- The digitizing and archiving process which will handle the digitization of the collection documents.
- The cataloguing process where the metadata entry, review and final quality assurance are done.
- Finally, presenting the collection online process in which the collection is made available for search and preview through a front end website.

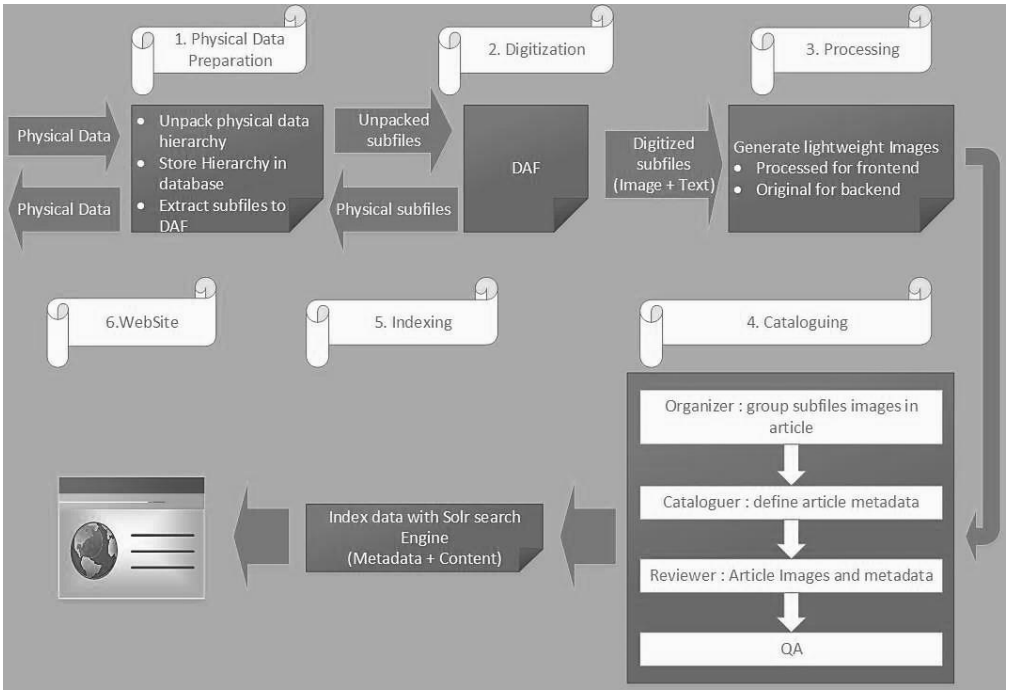

Figure 1. Digitization Framework Workflow

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

Indexing Process

The indexing process is the document entry point into the workflow where the collection is categorized and indexed into several hierarchical levels. The collection can have multiple hierarchical level according to the nature of the collection. The last hierarchical level is conceded the entry point for the next process, the digitization process as a digitization job which is labelled with generated

serialized barcode based on a pre-defined naming convention to facilitates and accelerates the document identification and tracking throughout the workflow by using a barcode reader

Digitization Process

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. DAF integrates with the tools for scanning, image processing and Optical Character Recognition (OCR) to assist repository administrators in managing the workflow.
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. BA provides DAF to the community as an open source tool (http://wiki.bibalex.org/DAFWiki).

Cataloguing Process

The cataloguing process is divided into three steps:

Step 1: Document definition step, where the digitized images of the document are grouped together using simple drag and drop technique. For concurrency handling, the sub file is locked by the user who started working on it until the user marks all the documents finished from this step.

Step 2: Metadata entry step where document title, source, author, page count and the rest of the defined attributes which are defined for collection type at hand are defined. The keyword linking is also done in this process for enhancing the search process.

Step 3: Metadata review step where the documents metadata is reviewed and corrected, then the document is marked "Finished". At this point, the cataloguing process of the document is complete. (See Figure 2)

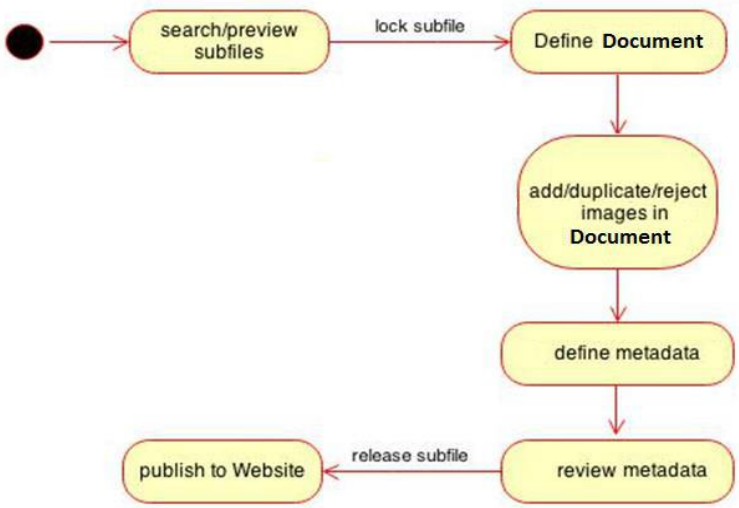

Figure 2. Cataloguing Process

Presenting the digital archive online

Publishing the digital archive is consider a challenge for large collection, as this requires offering the archive to the public whom are vary between normal users and research specialists with easy navigation and more advanced search options while maintaining high search results retrieval speed which is one of the major challenges, therefore the need for high speed search engine is vital for responsive front end. In the following subsection, the main components of the BA digitization framework publishing process will be explained in details.

Building search engine

BA digitization framework selected search engine is Apache Solr Search Engine which is based on Apache Lucene project v4.1. Solr is highly reliable, scalable and fault tolerant, providing distributed indexing, replication and load-balanced querying, automated failover and recovery, centralized configuration and more. Solr powers the search and navigation features of many of the world's largest internet sites. SolrNet API is used to connect to Solr server for fast retrieval of the searching results and the reason that Solr is selected that it provides many advanced features; such as: simple/advanced search, results highlighting, fields AutoComplete and morphological text search
BA implemented generic indexing application in which only the schema is changing according to the metadata fields of each collection.

Designing navigation dimensions

For each collection, one or more major dimensions, that control the data flow of this collection, can be extracted. Based on this concept, the navigation of the digital collection will easily direct the normal users through the collection. BA designed a generic front end website for any collection with any number of main dimensions to be used. The front end also using cross-referencing search to narrow the search results retrieved.

The Generic Document Viewer

BA developed a generic document viewer that is used for previewing any type of documents. The viewer architecture is based on two components; Server Side: RESTful services and Client Side: JavaScript using JSONP. The viewer support many powerful features which are image preview, metadata preview, text selection, searching/highlighting and zooming options: fit width/height. The viewer is based on several web services which are:

- **Metadata Web Service:** Retrieve document metadata and return all available technical information of the document images such as (width, height, page count)
- **Content Web Service:** Retrieve the image of each single page in the document applying scaling to custom width and height responsively according to the displaying device. Also, returns the selected text based on the user highlighted area.
- **Search Web Service:** Perform the search using Solr engine APIs in the content of the documents which has been through the OCR'ing phase. Also, highlight the matching phrases in the document image.

Case Study: CEDEJ Press Collection

Collection Overview

CEDEJ press collection consists of more than 800,000 pages of press articles dating back to 1976 covering one of the most critical periods in the history of Egypt. The collection spans over several subjects such as political, economic, social and cultural subjects from most of the Egyptian publishers. From the collection analysis, BA extracted the main metadata fields of the collection press articles. The standard press metadata fields defined are the title, date, publisher, authors, persons, keywords, subjects, organizations, countries, languages and the type (whether press or report).

Framework Implementation

Workflow overview

As shown in Figure 3, BA digitization framework designed the categorization levels of press collections to be organized just as the physical structure of the collection which is boxes each consist of several folders and each folder is divided into several files and each file has several sub files. The sub file represents a digitization job that is given a barcode label and ready for the digitization process. The indexing unit generates a serialized barcode based on a pre-defined naming convention

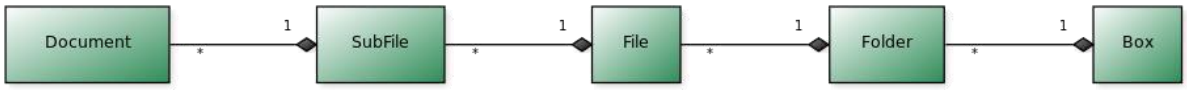

Figure 3. Classification Unit Workflow

Afterwards, the labelled sub file enters the digitization process and managed by DAF to produce scanned, processed and OCR'ed images of the sub file documents. Also, a light version of the document images are generated to be used in the cataloguing back-end. Then, the cataloguing process starts using the cataloguing back-end for grouping the images of each article.

The initial metadata are entered for each defined article for all the defined metadata fields as well as linking the documents keywords. After that the quality controller reviews the document metadata for approving or correction (See figure 4).

Figure 4. Cataloguing Back-end – Entering Metadata 6

After this step is done and the sub file is marked finished QA, The publishing process starts with accumulative indexing of the entered metadata along with the OCR'ed content into Solr search engine. Now, this document is available through the CEDEJ online front end website.

CEDEJ Online Archive (http://cedej.bibalex.org)

CEDEJ digital press archive is now online with the progress of more than 115,000 press clips finished so far varying between Press and Reports with more than 200 publishers, more than 14,000 writers and reporters and over 100 different subjects. CEDEJ online archive offers the navigation through four main dimensions: Timeline, Authors, Publishers and Subjects.
CEDEJ online archive offers advanced search with cross-referencing facility which make it easier and faster for both public and research users.

Conclusions

The role of Bibliotheca Alexandrina in the establishment of the CEDEJ Press Collection Digital Archive, using the proposed digitization framework, 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 digitization framework has been proved to be a configurable and flexible management framework for any digitization workflow.