

The Road to Providing Access to Kenya's Information Heritage

Digitization project in the Kenya National Archives and Documentation Service (KNADS)

Francis G. Mwangi

Kenya National Archives

Abstract

This paper discusses how the Kenya National Archives and Documentation Service (KNADS) started digitizing its records in its quest to give access to millions of documents in its holdings. KNADS under the law has the responsibility of taking “all practicable steps for the proper housing, control and preservation of all public archives and public records in Kenya”. (Cap 19 Laws of Kenya) The paper indicates that the purpose of the programme is twofold, to give access to the information contained therein and to preserve the original archival materials for posterity. The paper shows the work plan that KNADS adopted in digitizing a part of its collection; it shows the different methodologies adopted by KNADS to achieve these objectives, and the challenges it has faced so far. It concludes by indicating that although considerable progress has been achieved by KNADS in ensuring that the most consulted records in its collection have been digitalized, there is a need to look for new methods of achieving its goal in a shorter period as well as giving access to those records it has digitalized.

Author

Francis G. Mwangi was born in Kenya in 1967; currently he is the acting Deputy Director of the Kenya National Archives and Documentation Service, Administration and Finance. He is also the head of the Films and Audiovisual Records in the Kenya National Archives. He is a holder of a Master of Science Degree in Archives and Records Management from Kenyatta University, Nairobi.

1. Introduction

KNADS is a Department in the Ministry of State for National Heritage and Culture in Kenya. It was established by an Act of Parliament in 1965 to take all practicable steps for the proper housing, control and preservation of all public archives and public records of enduring value, and make them available for public access (Public Archives and Documentation Service Act, Cap 19 Laws of Kenya).

The weight of this responsibility is rapidly changing in the current digital environment and has become a real challenge. It's worth noting that some of the archival records and media are old, brittle, and delicate that requires careful handling. It's therefore important that the Kenya National Archives and Documentation Service actively intervene to ensure that these records will be available today and in the future. The use of archives is the goal that all archivists would endeavour, but the availability of archives for use by the public, and indeed all other aspects of archives management depend on archives being properly preserved and cared for, and now being made available to the users all over the world through the techniques of today and therefore the need to digitise our records.

Traditionally archivists have shaped their preservation activities around the notion of Permanence; their objective has been to ensure the permanent preservation of archives. This is despite the fact that no record, no matter how well protected and cared for, enjoys an unlimited lifespan. Internal processes of decay ultimately defy even the most sophisticated intervention by archivists.

Since 2007, KNADS has carried out a digitisation programme. This has involved digitizing some of its oldest and heavily used archival materials, some dating back more than 100 years. This programme has resulted in the digitisation of close to 12,000,000 documents of archival records (KNA/8/2 Vol.11). While this may sound a big number, it is a very small portion (3%) of the more than 400,000,000 pages of archival materials in the custody of KNADS. It should also be noted that digitisation is both capital and labour intensive as will be shown by this paper.

2. Background

In 2007, the Kenya National Archives and Documentation Service (KNADS) decided to start digitizing the records of Coast province that are held at its Nairobi headquarters. They were estimated to number slightly over 1.7 million pages of both bound and loose pages. These records were selected because of their age and heavy usage by clients; it was argued that reformatting them into digital copies would allow wider use and ease of access while preserving the original. Constant handling of these records by users over the years has exposed them to wear and tear. Continued usage of the original document would lead to their destruction. By digitizing these records, the original would safely be preserved while the digital surrogates would be used for access. By virtue of them being digital, it would be possible to produce surrogate, and derivative files without any damage to the original digital master. The digital object would then retain all significant information contained in the original document(s), and under appropriately stringent conditions related to migration, refreshing, and backing-up of the original file, it should survive over time.

2.1 Ease of Access

Demand for access to original materials, often termed as ‘primary source materials’ is increasing every day. Members of the public often require instant access to records. Archival records are mostly single copies and as such multiple users cannot access them at the same time. However, with digitisation, it will be possible to ensure these records are available to the users online where multiple users can access them. The trend worldwide in archival institutions has been to digitise and even make records available on the internet where users can access them after paying a stipulated fee.

It is envisaged that the digitisation will be undertaken to create a very high quality archival-master level quality that will allow for multiple output (e.g. print, microfilm, access images, thumbnails, etc.) when need arises.

2.2 Baseline Assessment

A baseline assessment was carried out by the department to establish the nature of the source materials. About 30% of the records were found to be very fragile and brittle and therefore, they required special handling and equipment when scanning in order to avoid any damage. About 10% of the documents required wide format scanners, as they were not of standard size. Big portions of the document were hand-written and therefore, Optical Character Recognition (OCR) would not be performed on them.

2.3 Project Outcomes

After these materials have fully been digitised, users will no longer need to access the original materials unless on very rare occasions to satisfy curiosity, for authenticity and legal purpose. It will also be possible to do full text searches on the records as they will be indexed and as such it will be easier to automatically identify documents containing relevant information, something that is not currently possible with the paper records.

3. Progress of Implementation

The first three years of the digitisation project was outsourced to private firms. This was due of the fact that the department did not have the internal capacity in terms of adequate skilled manpower and equipment. This period saw the digitisation of over 10 million pages of archival records. The contracted firm had a workforce of over 40 people working in two shifts (day and night). In the last two financial years, the department decided to undertake the project in-house with a view to developing internal capacity for sustainability. The department procured twenty computers, twenty medium duty scanners and one wide format scanner. The intention was to use the department's staff to undertake the exercise.

4. Roadmap to Digitization of the Records of Coast Province in the KNADS

4.1 Work Plan and Payments

Mile stone one: Digitization of Coast Province Documents			
OBJECTIVE	RESPONSIBILITY	MILESTONE/INDICATOR	BUDGET (KES)
Pre-Digitization Accession 11 th - 22 nd June 2007	Outsourced Company	1,685,000 pages unclipped, dusted, demagnetized and batched at KES. 0.50	842,500.00
Scanning 22 nd June – 10 th August 2007	Outsourced Company	1,685,000 pages digitally captured, OCR. edited and indexed, KES. 2.00 per document.	3,370,000.00
Post-Digitization Accession 22 nd June – 15 th August 2007	Outsourced Company	1,685,000 re – clipped, transported to storage and catalogued at KES. 0.25	421,250.00 100% payment on completion

4.2 Workflow

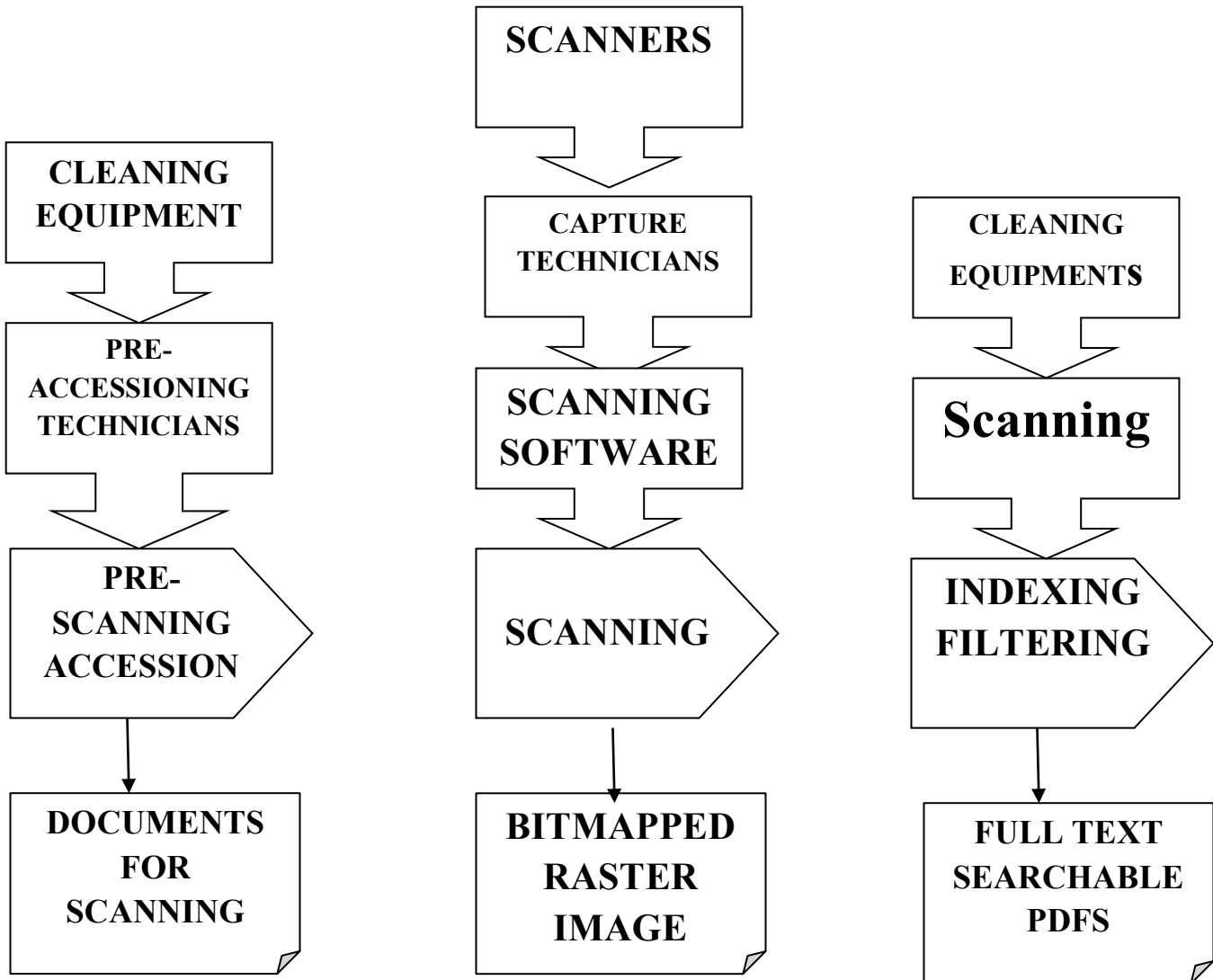


Figure 1. Workflow and events leading towards searchable document in KNADS

5. Methodology

The first step was the preparation of the 1,685,000 documents through cleaning. Demagnetizing, unclipping and batching that mirror the output indexing and metadata.

The second step was scanning with various equipment according to documents specifications. All the original 1,685,000 documents were scanned into digital master documents in TIFF CCTT4 format (resolution 6000 pixel – 8 bit grayscale) with lossless compression. The raster files were then saved into a backup media repository.

The third step involved filtering the documents into the required formats. Filtered copies from the digital masters were made according to the needs. PDFs were done with metadata and full-text search capabilities for retrieval done through OCR and documents schemas while GIF and JPEG thumbnails were created for intra/internet display. The work was divided into 5 objectives as shown below

5.1 Objective One: Pre-Digitization Accession

Input: 8 members of staff, vacuum cleaners and demagnetising hardware

Activities: 1. De-magnetising the documents
2. Dusting
3. Unclipping documents
4. Batching

Output: 1,685,000 clean unclipped and batched documents

Timeframe: Two weeks

5.2 Objective Two: Scanning

Input: 4 capture/editing technicians, scanners, computers and softwares

Activities: 1. Scanning
2. Editing
3. indexing

Output: 1,685,000 indexed raster images

Timeframe: Eight weeks

5.3 Objective Three: Indexing and filtering

Input: 6 capturing/editing technicians and sorting staff

Activities: 1. Optical Character Recognition
2. PDF Conversion
3. Document Schemas and Metadata tagging

Output: 1,685,000 text/metadata searchable PDFs

Timeframe: Four weeks

5.4 Objective Four: Post Digitization Accession

Input: 6 re-clipping casuals

Activities: 1. Re-clipping documents
2. Transporting Documents
3. Arranging/Cataloguing documents

Output: 1,685,000 catalogued and achieved hard copies

Timeframe: Two weeks

5.5 Objective Five: Installation of Hardware

Input: 2 Hardware & Software Engineers

Activities: 1. Installation of backup external disks
2. Installation of Server

Output: Configured Storage Hardware System

Timeframe: One week

5.6 Scanning Hardware Used

- Two high duplex scanners capable of scanning 80 pages per minute, but given that these are archival document, the output is much less than that.
- 4 book scanners one being planetary capture scanner
- One wide format scanner (A0 size)

Storage Requirement: Master Documents – TIFF: - 1685,000 pages x 2.4 megapixel = 3,840,000 megabytes = 3,840 gigabytes = 3.84 terabyte of space required

Filtered Compressed PDFs: 1,600,000 pages x 120kb

Software requirement: to manage the repository for retrieval, Adobe Acrobat and Acrobat Reader are being used.

5.7 Quality Assurance

A team of officers from KNADS were detailed to carry out quality control analysis at various stages of the project to verify that all reproduction is up to standard. The quality assurance analysis was carried out on a random sample of 10% at all of the stages. The quality assurance analyses were to determine the following:

- Size of image;
- Resolution of image;
- File format;
- Image mode(i.e. colour images are in colour, not grayscale);
- Bit depth;
- Details in highlights and in shadows;
- Tonal values;
- Brightness;
- Contrast (e.g. stark black and white contrast on anything except simple line drawings);
- Sharpness;
- Interference;
- Orientation;
- Noise;
- Alignment of colour channels;

- Cropped and border areas, missing texts, page numbers, etc.;
- Alignment of images;
- Missing lines or pixels;
- Text legibility and meta data capture.

6. Challenges

The following are some of the challenges faced during or immediately after finishing the scanning exercise.

- There were 1,585 wide format drawings and maps which had not been captured initially and which form part of the records of coast province. Since the drawings and maps are an integral part of the contents of the repository, these drawings and maps have to be scanned and according to the firm we had outsourced, the cost will be at KES. 150.00 per drawing/map, this was because the maps and drawings would require wide format scanners. After shopping around we found that the price was actually KES 50.00 less than the market price.
- We also realized that there was another consignment of 190 boxes in the basement repository with 437,000 extra records of the same provenance, i.e. Coast Province that had not been captured in the initial count as they were in another repository. The cost of digitizing them was charged as the earlier records.
- We also needed to store, make backups and manage the electronic repository created with two 500GB external hard disks, one hundred and twenty 4 GB DVDs and PDF archival software, all totalling to an extra KES. 581,500.00
- Quality: While the contractor has been given the specs for digitizing the records, where mass digitization is taking place, quality may be compromised, especially given the fact that archival quality digitization requires a high resolution. This meant that an archives staff member has to be a part of the process all of the time. The work has to be checked again and again to ensure the right image is always achieved.
- Expenses: As will be seen, there are always “other” things and issues that crop up during or after the completion of a project. Therefore, apart from the original budget, extra funds have to be kept aside for such issues and things that are bound to arise.
- Consistency of the filing system: It was noted that during quality control, it should be ensured that the filing order of the manual documents is maintained in the digital document, the documents in every file should be scanned from back to front just as the file has grown, where a document has more than one folio, the scanning should start from page one of the last page and in order to ensure that this was done properly, the departmental committee dealing with monitoring and evaluation of digitization programme was to closely and regularly visit the site and assess the progress.
- Re-boxing and re-shelving of files: This is a major challenge during a digitization exercise especially where the services have been outsourced. The archives developed a stamp having the inscription ‘verified’, ‘date’ and ‘signature’ and the re-shelving team had to look and ensure that the records had been digitized before they are returned to their boxes.
- Outsourcing: While it is appreciated that outsourcing resulted in the digitisation of over 10 million documents within three years, the exercise was very expensive and yet there was no

specific budgetary allocation for the project. This meant that the project could not be sustained through outsourcing.

- **In-house Project:** The initial intention was to deploy officers from other sections to carry out the digitisation on a rotational basis. The officers could not afford to dedicate adequate time for the exercise since they already had other duties to take care of in their respective sections. The department, therefore, resorted to hiring of casuals to assist staff in the digitisation exercise. The money allocated during the last two financial years for hiring of casuals is a cumulative total of 3 million shillings. This figure has only been adequate to hire 20 casuals for a total 9 months. This implies that half of the year no digitisation takes place yet the total number of people undertaking the exercise is only half what the external contractors used to have. Furthermore, considering that the programme is long-term, it is also not sustainable to use casuals year-in year-out.
- **Access:** Although the primary reason for digitization was to offer access and preservation of the original, the actualization has been slow. This is because of the following reasons:
 - **Website:** The Kenya National Archives and Documentation Service website had been hacked and it took time constructing another one.
 - **Content:** Hosting the website with all these data became an issue, the capacity of most of our service providers was low and as such they could not handle our request.
 - **Technological Challenges:** There was no one who seemed to know exactly how to go about uploading the material that we wanted uploaded without compromising the rest of the material in the server.
 - **Payment and Charges:** If we were to charge for the material accessed, how would we go about it and what modes of payment would be accepted, noting that this is a government department.

7. Way Forward

Long-term solutions to the digitisation project needs to be found with a view of ensuring that it is sustainable. These would include the following:

7.1 Personnel

To clear the huge backlog of 97% of un-digitised materials, the department would require engaging 200 staff on a full time basis. That number of staff would be able to digitize 200,000 documents per day at the rate of 1000 per person. At that rate, the department would be able to clear the backlog in five years.

7.2 Training

Digitisation is a technical area that requires professional skills to ensure quality output. It is therefore important for all the staff engaged in the project to undergo thorough training in digitisation both at middle and advanced levels.

7.3 Equipment

While the department has already acquired digitisation equipment, these would not be adequate if the department was to acquire the desired number of staff. It would therefore be necessary to acquire more equipment. Additionally, ICT equipment becomes obsolete very fast. It is therefore important to budget for the upgrading and replacement of the equipment that is no longer serviceable.

7.4 Budget

As noted earlier, digitisation is a capital intensive project and as such the department needs to introduce a specific budgetary item for this purpose with adequate allocation to cater to its needs.

7.5 Partnership

Another way of going about the digitization programme is to partner with stakeholders and friends of archives wherever they may be, the partnership can be in form of assistance, technical knowhow, equipment, financial or any other way that can be offered to make the programme a success story.

8. Conclusion

Since archival material exist in single copies, they had to be handled with utmost care. For this reason automatic document feeders were avoided, unless the papers were in very good physical conditions and of standard size. Another rider adopted was the re-assembling of all documents into their respective files and the cover of the file stamped 'digitized'. Finally, since the files are read from back to front, digitization had to be done in the same way with folio one starting off till the last folio to be filed becomes the last.

Though it had been hoped that by undertaking the exercise in-house the challenge of sustainability would be resolved, this has not been forthcoming as the department is too thin on the ground in terms of personnel. Hiring of casuals was an appropriate stopgap measure but it is also not sustainable in the long run. It is, however, worth noting that the department now has adequate equipment for digitizing records. However, the challenges of internal capacity need to be resolved with long-term solutions that would make the project sustainable. The international partners and friends are invited to assist the archive move a step further than where we are by making what we have already digitized available to the world.