1.1. Various Stages involved in Digitisation Process

The various stages involved in the Digitisation Process with reference to the Generic Methodology and Digitisation workflow are described in detail hereinunder:

STAGE 1 Definition of the project:

> The Digitisation project is initiated by collecting the client requirements and preparing a detailed specification document based on these requirements.

STAGE 2 Definition and creation of workflow:

> The workflow is engineered to monitor the digitisation process to adhere to the defined specifications.

STAGE 3 Preparation of Digitisation facility:

> This stage involves the deployment of Scanners, Computers, Software, install Network and Personnel on site

- Site is inspected for the requisite infrastructure i.e Power, space, air conditioning, etc
- Chairs, Tables, racks and other furniture requirements and shop floor layout determined.
- Locations of the servers, switches, Scanners, computers, stocking areas, prescanning areas, processing areas and finishing locations determined.
- All hardware and software required are preinstalled and brought to site.
- Networking of the PCs done.

Deployment of Manpower

- The manpower required for the work is identified according to the roles envisaged in organisation chart.

- Each of the roles are handed over their responsibilities and the tasks that they need to follow.

  The managers are briefed about the work flow/process and they in turn would train up their respective wards.

STAGE 4 Identification of media to be digitised:

This stage describes the process of collection of books from the Client, segregating them, bar-coding and preparation of books making them ready for scanning.

- Center-In charge shall go to record bank and receive the documents.
- A complete inventory management mechanism would be maintained to track the documents and ascertain the various stages that the documents would be in.
- Following details will be gathered when documents are received.
    - § Title of the books received
    - § Condition of the books received (damaged, pages cut etc.,)
    - § Place from where it has been brought
    - § Name of the person who received books
    - § Date of the books received
    - § Time of the books received
    - § No of pages in each of the books
    - § Name of the person who issued the book

- § Acknowledgement will be given to concerned official for the books received by Center – In charge with details of Receiving person, Received date, Received Time, Record Title and Record Condition
- Similar procedure shall be adopted for issuing the books to the Managers/Shift In charges and from Shift in charge to the operators.
- A control Sheet is created for each book, which will travel, with the book across the workflow. All documents/books that are collected are entered into the DigiFloand a physical document track maintained

STAGE 5 Transport media for Digitisation:

The media to be digitised is transported and made available at the digitisation facility.

STAGE 6 Categorisation of Books for metadata and Indexing prior to digitisation:

- Books are cleaned, coded and any tags in the books are removed prior to being issued for scanning.
- Any restoration to be done on the book/document will be done after setting a standard qualified by the client.
- Each book will be given a unique id (barcode/code), which is auto generated and the book is also identified with the OCLC number that exists in the books XML data (MARC21 record) that is downloaded prior to the Metadata capture process. This OCLC number and the auto-generated barcode will form the basis for tracking of the book across the scanning process.

STAGE 7 Digitisation of media through workflow:

- The books that have been cleaned, Meta data entered and documented are passed to the scanning operators through the DigiFlo. This ensures
  - No duplication of the scanning
  - A clear idea of what stage of scanning/processing a record/document is in.
  - Effective planning of the scanning process.

- o Time stamps of the issue and the scan are available.
- o Details of the scanning operator are available for analysis of performance of each of the operators/teams

Scanning done through the workflow ensures reliability and consistency in quality of scanning. Scanner parameters sent through the workflow ensure consistency. All encountered errors in documents are trapped during scanning.

- Books will be assigned to operator for scanning by Team Leader
- Operator will scan the books while viewing the details about record in the screen and enter remarks about record like Blank Page, Damaged paper etc.,
- Before commencing the scanning the work flow automatically sets the following parameters
  - o **Resolution** :   300 dpi or above and fix based on specifications
  - o **Compression**:   As required
  - o **Brightness** :   As required
  - o **Contrast**   :   As required
  - o **Background** :   White
  - o **Colour**   :   Black & white/Grey Scale, 24 Bit RGB
  - o **Size**   :   Set as per the size

- Directory structures will be created when scanning starts and start date and time fields will be updated with system date and time. Once the user finishes scanning, he will press completed button, then scanned pages in TIFF will be saved into concerned directory/folder and end date and time will be updated with system date and time.

STAGE 8 First-level Quality check:

After scanning , the books undergo a quality check for the verification of quality of scanning, resolution & clarity which is both automated and through eyeball check.

This is the first level of quality checking which ensures that any defects are not carried across the entire work flow. Errors trapped at this stage make sure that these are not carried forward to subsequent stages resulting in unnecessary redoing of work.

The Quality control at this stage will be of two kinds

- Automated QC Check through the Document format Auditor of TIL it checks for
  o Verifying the Color / Black and White
  o Compression
  o Resolution
  o Page Size (Size set while scanning record)
  o Comparing Number of pages scanned with scanned TIFF Images (verify if all pages are scanned)
  o Visibility of the page i.e. page scanned is clear or not, any noise exist.
  o Verifying Page Quality
  This ascertains whether the scanned image can go to next stage.

- The second part of this quality check is the eyeball checking this helps us inspect for
  o Damaged pages
  o Missing pages
  o Flying text
  o Edges of text cut

Once the quality check is completed the books that will have to be re-scanned and to be sent back to the scanning operator for re scanning would be identified and the procedure is repeated on the rescanned pages.

STAGE 9 Processing:

Processing involves the cleaning of documents and making them suitable for subsequent automated processing stage. The operations done at the preprocess stage are:

- Cropping to remove peripheral noise
- Any large noise bands and scratches are removed
- If the documents are scanned in a spread mode page splitting is done.
- Pages with images and half text are segregated as they would require different automated settings
- Left and right pages are segregated.

On completion of the preprocessing the documents/books are subject to the automated processing following is done in the processing.

- Deskewing
- Despecking
- Noise removal
- Resizing
- Padding
- Uniform sizing of all pages
- Rotating
- Brightness/Sharpening
- Background.
- Growing /Smoothening

The Processed books are then subjected to the next quality check.

STAGE 10 Quality Check:

QC1 to check on the quality of the Procesed Tiff documents. Eyeball and Automated Checks for parametric and subjective Quality.

Quality Check 1 is done to ascertain whether the parameters that have been set for the documents have not been tampered with during the processing

stage and that the documents have the necessary levels of quality both subjectively and objectively.

Automated Check: This is a check which automatically checks for the basic parameters of the documents and throws out exceptions which exist which do not conform to the specification. This help us to handle the exceptions rather than the whole document. Following are the parameters that are checked. The present version of Workflow has a provision

- Verifying the Color / Black and White
- Compression
- Resolution
- Page Size
- Number of pages that exist against the number of pages in the meta data
- Blank Pages

An eyeball check is done and checked for the following parameters

- Amount of skew
- Noise levels
- Any pages with text edges cut
- Readability of the document
- Background Color
- Proper padding
- Missing pages

The errors if any are logged through the DigiFlow and the same is saved as a error log in a text format and sent for reprocessing.

STAGE 11   Indexing / Creation of metadata:

The indexing and creation of metadata for Structured data is done as shown in the flow diagram for corresponding types of media.

As the documents are of various types and the categorisation and segregation of the documents is essential two type of profiling is done based on the following parameters

Meta data eventually forms the core of any digitisation process and this is the ultimate seek item for a document. Meta data could be printed and verfied manually this would involve eyeballing of the documents and checking on the spelling and its veracity.

DigiFlo interface is created which will capture the relevant metadata from the downloaded MARC21 XML data file and populate the database for that book. This process is done through the Importing data from XML data file tool whose operation is explained in the DigiTool Chapter Pt. 11.11.

STAGE 12   Conversion / Rendering to required output formats:

The documents are now converted into any format (e.g.: PDF, MSword, Text) and the attributes of the document are set depending on the type of document.

Based on the profile of the documents the following permissions are set

- Save
- Copy
- Print
- Password
- Content Change
- Commenting
- Other attributes

STAGE 13   Backup and archiving :

These operations are done as per the client provided specifications and formats according to the weekly/daily schedules,Preparation of CD/DVD for content delivery.

STAGE 14 Return of print Media:

The books are reorganised to be returned to the client at this stage.

On completion of the Quality control the physical documents and books are

- reorganised
- checked for the sequence
- handed over to the client
- document control register is created
- scanning and the meta details of the record are printed out and attached to the document.
- Acknowlegement is taken from the client

STAGE 15 Digital Deliverables :

The data is delivered according to client stipulated standards. While handing over the digital media to the client is done the following is taken into consideration

- File and directory structure
- File and directory naming conventions
- File formats
- Resolutions
- Compressions
- Permissions
- Meta data in XML/Excel/Data base formats
- Error lists and document exceptions
- Administrative meta data
- Structural meta data

The parameters mentioned will be checked through Thrinaina's Document format Auditor which generates the reports in the requisite fashion.

STAGE 16 Uploading the delivery and archiving (if required by the client):

The data is uploaded to the database server and the archival/retrieval application is hosted on the application server. The final delivery is then uploaded to the database server is linked with the archiving application hosted on application server. The books can be searched online through Internet or over the intranet.