10. Digital Library @C-DAC Kolkata

Dr. Amiya Baran Saha & Utpal Kumar Saha, CDAC, Kolkata

Introduction

Centre for Development of Advanced Computing, Kolkata, one of the four Mega Regional Scan Centres is working in the Eastern part of India for the digitization of library documents like books, manuscripts, magazines, journals etc, under the Digital Library of India Project sponsored by Department of Information Technology, Ministry of Communication and Information Technology, Govt. of India. Besides book digitization works, C-DAC, Kolkata is also developing Cross-lingual Information Access, Multi lingual Crawler, Bengali OCR with work flow, Multimedia Interface for physically challenged in Bengali, Automatic search Indexing tools, Multilingual and Multimodal authoring tools, Text summarization in Bengali, Folk songs search and retrieval etc.

Steps of the Book-Digitization

To achieve the best quality of Book-Digitization work, care should be taken at every step of processing starting, from the selection of books (for avoiding copyright act) to the Metadata preparation, Scanning, Cropping, Cleaning and OCRing including storing in DVD for web enabling.

Metadata:

Metadata accompanies and makes reference to each digital document and provides associated descriptive, structural, administrative, rights management, and other kinds of information. This metadata will also be maintained and will be migrated from format to format and standard to standard, independently of the base document it describes.

Three types of Metadata creation are generally followed in the Digitization Process, viz,

- REGULAR Metadata
- ADMIN Metadata
- STRUCTURED Metadata

One has to be prepared before the scanning process and the other two are to be prepared after the scanning as follows: -

Pre-Scanning Process:

- Regular Metadata Information is created for all the Documents (books, periodicals, manuscripts, etc.) available for Scanning. It needs the necessary Fields of Information like Title of the Document, Author/ Creator, Year of Publishing, Publisher, Subject, Language, Source Library, Total Pages, and Type of the Document.
- This Regular Metadata is stored in the Database as .xml Format.
- The Extensible Markup Language (XML) is a data format for structured document interchange on the Web. XML is intended to be a standard "to make it easy to define document types, easy to author and manage documents, and easy to transmit and share them across the Web." The extensible character of XML will also make the encoding of metadata easier and more flexible.

Post-Scanning Process:

- After the Image-Processing and OCRing, it needs to prepare the other two types of Metadata.

  a. Admin Metadata is used for administrative purpose. The necessary Fields of Information are Barcode (auto-generated), Operator, Scanner No, Scanner Date, Vendor Name, Scan Center Name, Digital Re-Publisher, Digital Publication Date, Numbered Pages, and Unnumbered pages.

  b. Structural Metadata is used for accurate and pin-pointed Information Search and Retrieval from the Full-Text Database. The necessary Fields of Information are:-

  - Barcode (auto-generated)
  - No. of Chapters
  - Chapter Details
  - Name

  Related TIFF Location and Details

  - Associated Keywords
  - Keyword 1
  - Keyword 2

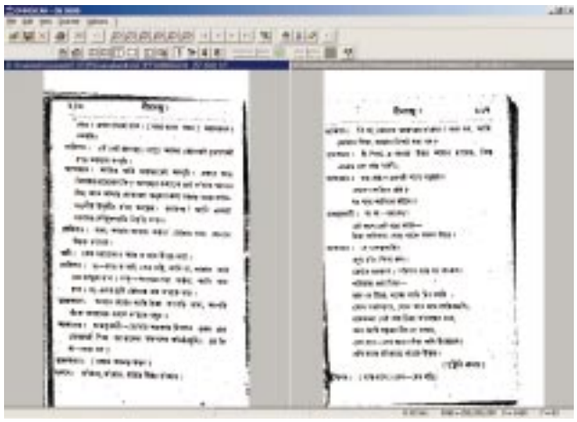
- The Admin Metadata and Structural Metadata are entered and submitted along with the Scanned Documents

Scanning Process:

Due to ageing of the documents the physical quality of the papers is being transformed into yellowish or brownish in color and also become brittle in nature. Those documents are needed to handle with care during the scanning job. In earlier days, libraries are used to preserve their old books and documents by laminating the pages with tissue papers on both sides for lasting. To digitize such type of documents, there are a few parameters of Scanner like aperture, brightness and threshold should be kept in mind during the scanning process. Otherwise the scanned image will be either darker or lighter, which is not acceptable quality of scanned document. Here are typical examples for Zeutschel OS 5000 Scanners to illustrate how to set the values for those parameters for different qualities of papers.

Quality of paper	Aperture	Brightness	Threshold
Normal paper	5.6-8	5-6	75-90
Laminated with tissue paper	5.6-8	3-5	100-125
Brownish or Yellowish paper	4-5.6	6-8	90-100

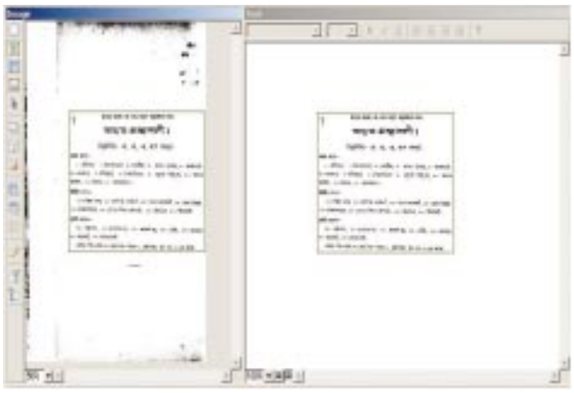
The scanned images are to be saved in .tif extension with a resolution of 600 dpi. The text content is to be scanned in Black & White (1 bit) mode while the images are to be saved in gray scales (8 bit) mode for best reproduction in Digital Library work.



A typical Scanned image page of a book, laminated with tissue paper

Cropping:

The scanned images are needed to crop for further processing. *ABBYY FineReader 7.0 Professional* is one of the useful software for cropping of the images. All the images are to be added in a batch. Then the images are to be analyzed into blocks. Blocks can be categorized in three different types like Text, Table and Picture. The Analyze utility automatically recognizes the respective blocks according to the content of the images and it draws the blocks very closer to the content. The blocks are to be checked for all the images if correction is needed. Finally the images are to be saved with blocks content in every image with the source file name. To make the operation faster a Function Key enabled tool is used with some macro applied on it. The advantage of using this tool is to recognize the block type and save the images in a batch process automatically. The main disadvantage of the analyze tool is automatic recognition of the block, because there is a chance of leaving of lighter scanned document. Therefore all the images have to be checked whether any part of the content is kept outside of the block or not.



The block of the content will be generated automatically according to the area of the content.

Another useful *Image Cropper* Software is also used for cropping of the images. Here a block is to be drawn manually around the content of the image but the cropped image will be saved automatically with the source file name after cropping the image. There is no batch process utility to crop the images. The cropping of the images is normally be done by one after another. The advantage of this tool is the manually recognition of the content area, so there is

no chance of content loss from the images. The only disadvantage is the more time consuming due manual selection of the area to be cropped for all the images one after another.

The block of the content is to be drawn manually according to the area of the content.

Cleaning & Resizing:

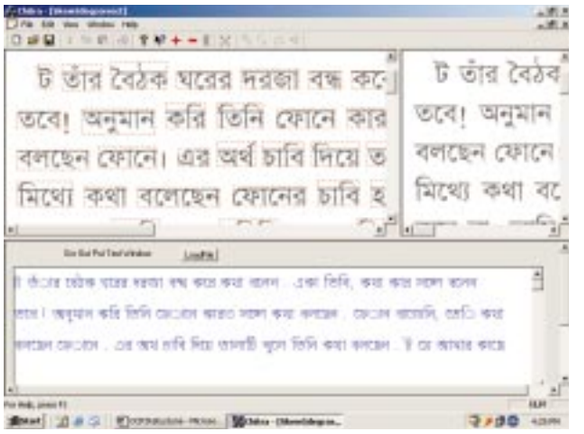
A popular tool used for cleaning and resizing of the images *is ScanFix 4.21*. There are a few parameters for smooth operation of this tool. The most commonly used parameters are De-Skew, Noise Removal, Resize, Smoothing & Completion, Intelligent Crop, Manual crop & pad and String Options. These parameters are to be set with some standard / appropriate value according to the requirements. This process can be done with the help of Batch process utility. It saves the images with source file name in the user defined target directory automatically. This tool makes the operation faster by processing and saving automatically for all the images of a document applying all the parameters specified by the user. During this process some documents may be hampered from the content point of view due to the wrong parameters setting like Noise Removal, Intelligent Crop and Smoothing & Completion. Care should be taken so that the value of the parameters should not cross the maximum limit.

A typical skewed and noisy image cleaned & resized by using this tool.

Chitra – The "Bangla" OCR:

Windows based Bangla OCR is presently under development for Bengali text document. It is having the following features:

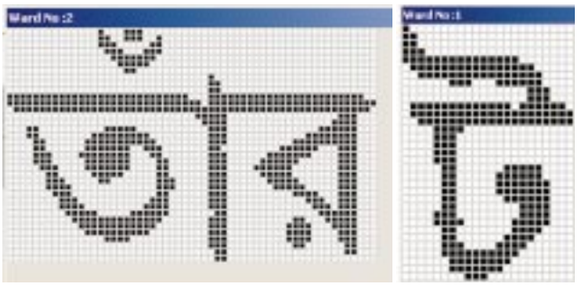
- The application is designed based on windows Multiple Document View architecture. The GUI based Chitra Software is engineered for ease of use. The view is having three-way splitter window as common in most of the OCR application.
- The Image can be acquired from the stored file or directly from the scanner.
- The image preprocessing functionality such as skew correction, gray scale to binary conversion is a part of the application.
- The provision of saving the scanned or processed image is also possible.



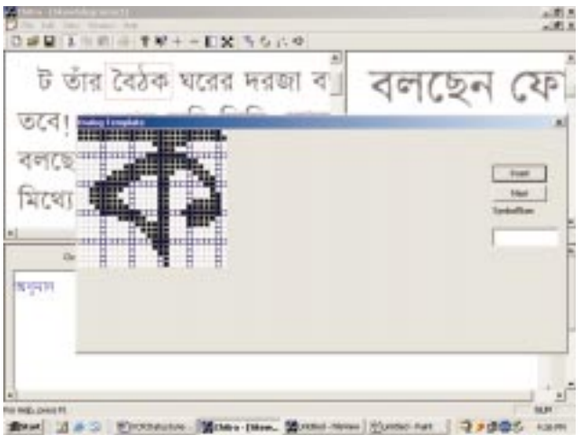
Main application window of Chitra.

- The GUI supports an OCR conversion of any small portion or of the whole document.
- The word level segmentation and character level segmentation viewer is available.
- The GUI Template file browser helps in addition/ deletion of the Template file.

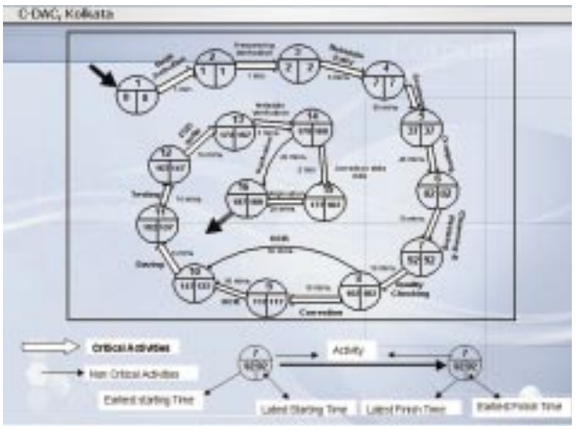
Multiple templates for different font type is also can be accommodated.



A typical Word level segmentation & A typical Character level segmentation.



A Template addition



A typical PERT – CPM Chart for Digitization of a Typical 250 Page Book

Conclusion:

Knowledge (i.e. Vidya) is the only commodity that grows, instead of diminishing, on distributing freely. It is wished that with the above humble and sincere effort of C-DAC, Kolkata, there will be additions, (however meager!) to the world corpus of available knowledge.