Case Studies

Digitisation Projects of

CASE STUDY 3: Million Book Project

The Million Book Project, led by Carnegie Mellon University School of Computer Science and University Libraries, aims to digitise a million books and beyond. Working with Government and research partners in India and China, the project is scanning books in many languages, using OCR to enable full text searching and providing free-to-read access to the books on the web.

The National Science Foundation (NSF) awarded Carnegie Mellon $3.63M over four years for equipment and administrative travel for the Million Book Project. India is providing $25M annually to support language translation research projects. The Ministry of Education in China is providing $8.46M over three years. The Internet Archive has provided equipment, staff and money. The University of California Libraries at Merced funded the work to acquire copyright permission from U.S. Publishers.

India, China and the U.S. agreed in November 2005 to join the Open Content Alliance (OCA), initiated by Brewster Kahle and the Internet Archive, because the goals of the OCA are consistent with those of the Million Book Project and the Universal Digital Library.

The Bill of rights for the information society articulated by Jaime Carbonell, Director of the language Technologies Institute at Carnegie Mellon, captures the vision of the Universal Digital Library: “Getting the right information to the right people, in the right timeframe, in the right language, with the right granularity”.

Requirement:

The Million Book project is designed to provide a wide array of content. Significant research is planned on the project, including OCR for Indian and Arabic languages and scripts. The research also includes developments in machine translation, automatic summarization, image processing, large-scale database management, user interface design and strategies for acquiring copyright permission at an affordable cost. For this to be achieved there necessitated setting up of scanning centers in India, China, Egypt, Hawaii and Carnegie Mellon.

The work handled at these scanning centers include design, procurement of Scanners, providing Computers, providing tools to process the scanned pages and deliver the same as per the Specifications designed.

To achieve all of the above and some The Million Book Project required Professional and capable organisations to Man and Operate these Scanning Centers. Thrinaina was successful in proving its capability to operate such huge operations and had since been allotted to operate one such Mega-Scanning Center.

Solution:

Thrinaina had made provisions for the capture of the overall needs and requirements of the Client through its detailed documentation and questionnaire. These capture all the details pertaining to:

1. Physical document details
2. Digitisation specifications
3. Indexing and Tagging specifications
4. Rendering Specifications
5. File and Folder structures/nomenclature

Based on the requirements specified by the client Thrinaina generates a clear and comprehensive specifications document incorporating all phases from Scanning to delivery as per the Client's requirement. The document also reports the number of physical pages that are mutilated and other pertinent information (in case of old Journals). This document will then be submitted to the client for approval and signing off.

The approved specifications along with the Physical documents will be sent to the Digitisation centre. The Digitisation centre is equipped with State-of-the-art Scanners, Software Computer systems installed with Thrinaina's Digiflo application. Screens will then be configured to capture the Meta and Structural data of the Journals. Thrinaina's Digiflo is then configured for the process with all of its predefined specifications. The Digiflo then tracks the progress of any individual document through its various stages such as Scanning, Cropping, and Processing. Reporting adherence and non-adherence to set specifications, any discrepancies with the specifications at all stages will filter out through the built in Quality Check modules of the Digiflo Suite. All deviations are corrected before being made available to the next stage of operation. This ensures that errors do not move forward in the flow but are corrected.

Once the images are processed to achieve clients specifications they are OCRd to achieve full text search ability (only in the case of English language books). The completed books in all respects are then delivered to the client to be uploaded and available for public usage through their Portal.

Learning:

1. Books which are approved for scanning are very fragile due to various factors relating to storage in the libraries, age, etc., have to more necessarily be scanned through a non-contact process using Book scanners. This protects the Books from further deterioration during the scanning process.
2. Capture of Metadata in formats consistent with the International standards is required as Digital Data generated through such Digitisation objectives have to be universally accepted. This ensures that duplication of work is absent.
3. Care should be taken while processing the books so as not to loose any information. This being the sole purpose of digitisation.
4. Reports generated should be able to clearly track the documents over various stages of operations. This is what Digiflo Suite with its accompaniment of Digitools ensures through its various MIS reports.
5. Tasks that involve substantial volumes require tools not only to process data but also track data. This has led TIL to fine tune its MIS and also include many new reports.
6. Confirming with delivery schedules is a must in such projects covering such wide geographical locations to avoid delays in subsequent deliveries through other processes.