Content Development for Digital Library of India

Lakshmi Pratha, Vamshi Ambati, Pramod Sankar
Regional Mega Scanning Center
International Institute of Information Technology
Hyderabad, India.

1 Introduction

In recent years, the availability of digital media for storage and exchange of information has increased in leaps and bounds. It has enabled us to communicate data, information and knowledge across the globe in a easy and efficient way. The availability of bandwidth to a large section of the society has also affected the way people access and use information in their lives. Libraries are known to be the most popular storehouses of information. Many communities in the last century or so, have developed rapidly in the presence of public libraries. These libraries host a large collection of books, journals, newspapers and other printed material which contains an enormous amount of knowledge.

However, these public libraries are centralised to their locations and are not readily accessible to everyone. There is also the danger of losing printed material in calamities like fire, floods and earthquakes. To overcome these shortcomings of public libraries, Digital Libraries are being built. A Digital Library will preserve the rare documents as well as a good portion of printed literature that would be of significant relevance to the society today. And this collection would be available for anybody, free-of-charge, 24 hours a day and 7 days a week.

In this direction, the Digital Library of India [5] aims at digitally preserving all the significant literary, artistic and scientific works of people and make it freely available to anyone, anytime, from any corner of the world, for education, study and appreciation of all our future generations. As the first phase of this project a million books would be digitized and made available on the web.

Given the enormity of the Digital Library of India project, its imperative that the job be distributed and decentralised over many Regional Mega Scanning Centres, scanning/processing centres across the country. Ever since its inception, then operating at 3 centres, the project has grown to 30 centres. The project has been successfully digitizing books, which are a delicate and fragile, yet dominant store of knowledge and culture. DLI now hosts close to one lakh books online.

The entire process of digitizing the book consists of various stages such as Procuring, Scanning, Image Processing, Quality check and Web Hosting. All these operations can be pipelined to improve the throughput. Another aspect is that these operations need not take place at the same venue. For eg. a book could be scanned at one place alloted only for scanning, Image processed at another centre which has good resources for image processing and then be quality checked at a Mega Centre and hosted there.

In such a highly distributed environment, where lot of data and control has to be distributed and shared, establishing a notion of collaborative effort and distributing discrete chunks of work becomes a high priority task. The streamlining of workflow needs to be automated to a large extent where information can be quickly and easily transferred from one centre to the other. As

part of DLI we work towards achieving a highly automated set up and the notion of a distributed environment which is flexible, yet cohesive yileding high quality output. In achieving this we have addressed challenges and have overcome problems and issues in almost all aspects of the digitization process.
In this article we present the various problems and challenges faced in the DLI and what solutions have been proposed and implemeted to overcome these. In section 2, we discuss the challenges in content development during the course of the digitization and web-enablement of books. In section 3 we describe the Process established in the project that helped us address these challenges and section 4 describes the Workflow of the digitisation process. Section 5 discusses the Architecture of DLI server and summarize in Section 6.

2 Challenges in Content Development

(a) (b)

Figure 1: Snap Shots of the Scanning (a) and Image Processing(b) Wings of Regional Mega Scanning Centre

The DLI Project is organized at national level with huge resources and man power. The DLI is the apex authority and it establishes, supports and monitors the various Regional Mega Scanning Centres. Each Mega Scanning centre has a number of scanning centres, processing centres and/or contractors under its purview. These sub-centres submit the digital content to the RMSC which hosts the digitized content on a webserver The DLI project has many practical constraints and challenges due to its distributed nature. All the challenges are centered around the data-centric architecture that needs stability and scalability [2]. In this section we present the various challenges in the content development for the DLI.

2.1 Massiveness of Data

Digitizing of a Million Books, is clearly a huge task, generating massive amounts of digital data. Storing, handling, maintenance and search and retrieval of this data is a huge challenge. The two important requirements of user satisfaction are quick turnaround time and high reliability. Meeting these goals while handling such huge amounts of data is a significant challenge.

Large Amount of Electronic Data In a book there are generally about 300 or more pages. A Million books would thus digitize atleast 300 million pages. When scanned and stored each

image of a page will be around 100 kilobytes of digital data. That means an average book will require atleast 30MB of storage space. However, the DLI stores not just the final processed images, but also the actual scanned ones as well, which would require another 30MB. When the OCR is performed the text is stored in different formats. These add up to close to another 30MB. Thus an average book would need about 100MB of disk space. Thus a million books would need 100,000GB. We can understand the massiveness of data considering that not many people would be using more than 20GB for their entire personal collection of music and data.
To ensure that digital data is not lost in hardware failures, natural calamities etc., multiple copies of the digital content is to be stored at different geographical locations and on different media, like hard disks and DVDs. This amounts to a mulitple of the 100,000GB that was calculated earlier. A DVD on an average can store about 50 books requiring about 20,000 DVDs. The physical storage of the DVDs is in itself a library management of its own.

Handling of the Electronic Data Apart from the storage of the books, DLI also maintains a number of details regarding each book in the form of metadata XML files. These are generated at various stages and provide significant information about the content of a book, the persons who were involved in the digitisation process and the quality of the final product. These files are also shared across centres for duplication checking etc. These files require to be stored in such a way that retireving information from them is quick and reliable. The files should be kept up-to-date with respect to other centres and multiple copies should be synchronised.
When a centre generates digital data, it has to be transfered to the Mega Centre its associated with. Transfering of such large amounts of data among centres is a challenge in itself. The maximum bandwidth available is just a fraction of the required, and thus data is transfered physically in HDDs or DVDs. The handling of this physical media, tracking and maintenance is a challenge faced by the Mega Centre.
The final product of the digitisation process is stored online on several terabyte servers. This content has to be available online almost always and should not crash. In case of a crash, we need mechanisms such that no content is lost and the server is back online without much unavailability of service. Also, the content should be searchable and the retrieval of data has to be very quick. Many of these challenges are too big for manual monitoring and tracking. Automation of several of these aspects is necessary for the project to function satisfactorily.

2.2 Challenges from Indian Language Content

There exists better support for English language than any of the Indian languages. English has a small alphabet and each of these alphabets are independent entities. In contrast, the Indian Languages have a large character set and many languages have *matras, samyuktakshars, shirorekha* etc. In addition, Indian language processing is considerably complex compared to English. This complicates a number of automation processes like Optical Character Recognition, search and indexing etc.

Representation and User Interfaces The representation of Indian language content is not unique. Initally, much of the Indian language content was stored in ISCII(Indian Standard Code for Information Interchange). To display the representaion of these ISCII characters, diferent vendors have developed many fonts. However, these fonts are not related to each other, and a page stored as a font code can be read and displayed only when the font is completely known. Recently, UNICODE has become a standard for Indian language content representation, but the shift completely to UNICODE shall take time. Many users also use ITRANS or its variant

OMTrans. It is a very convenient and popular transliteration scheme. It requires no browser or font support and the same ITrans text can be interpreted in any language and in any font.
To display the Indian language content, the data should be converted to the required font codes as per the user preference. The interfaces should also have knowledge of the language they are displaying and appropriately choose the fonts.

Lack of Robust Processing Schemes For Indian language content, the development of robust recognition systems is a major research challenge. Robust OCRs which give a high % of accuracy are still in the development phase and are not yet readily available. A number of language processing modules like the Morphological Analyser, chunkers and other modules like font converters and crawlers are being tested in real world scenarios. Search techinques, building indexes for search and other associated aspects are in the research phase and will take time to achieve expected standards.

2.3 Incomplete and Incorrect Metadata

Metadata is the data that represents different details regarding the book like the Title, Author, Publisher, Year of Publication etc. These details are useful when searching for books based on titile or those of a particular author etc. In DLI, metadata acts as an anchor of communication and coordinates the flow of data and information. Manual entry of the Metadata is error prone and the transliteration for Indian languages depends on the Metadata entry operator and is not standardised. This problem causes serious errors during Searching and Duplication detection.

Search and Retrieval The Search based on this erroneous meta data leads to confusion when, there was more than one copy of the same book, the search could return 20 different books or as book not available. The spelling errors on title, author and publisher, and the illogical and undefined dates and subjects etc. make the search ineffective.

Duplication The other major problem is the duplicate books being scanned or stored among and within each mega scanning centres. The number of duplicates increases as the number of books beign digitised increases, if the appropriate processes or tools are not used. There is another sub problem for duplication checking: retrieving the manually maintained records in many source libraries, archives and other locations which are in different languages across the country. The duplication of the resources can be identified only using the Meta information. If the metadata is incorrect, missing or incomplete it makes it difficult for determining the duplicates.

2.4 Distributedness of the Project

The entire project is being executed at various geographic locations, and the deliverable from each centre should be made avialable to the user seamlessly. The major problem is the collaboration point for making them available at one single point. All the books scanned at each mega scanning centre should be available for all other centres. The storage media used for storing the data needs a great deal of attention for preserving them on a long term basis. Thus the distributedness property leads to (a) Data Administration (b) Data Synchronization

Data Administration This process involves the handling of the data from the final stage of digitization till the web enablement process [4]. Maintenance of the data storage becomes the

key issue in the data administration. In case of using hard disks as the storage media, the hard disk crash is possible resulting in the effort of scanning, processing, OCRing precious books and resources being lost. This causes a huge rework to get the data back.

Data synchronization The data that has been stored at different scanning centres must be synchronized so that there is no duplication among centres. Also if one centre has lost its effort due to some reason the data unavailability on their server must be restored using the data form other servers.

3 The Process

Most of the challenges mentioned above have been addressed by adhering to a rigid Process and by establishing a decentralized Software architecture. The process at DLI is metadata centric.

3.1 Metadata

The metadata is represented in XML format and identifying the metadata that should be preserved along with the digital objects is a debatable topic. After several discussions at DLI the schema for metadata of a book, was decided as to comprise of three catgories as shown in Table 3.1

Category	Description	Usage
Regular Metadata	Title, Author, Date of publication, Publisher, Edition, Keywords, Subject, Language etc. (Dublin Core)	Search and Indexing
Admin Metadata	Library, Scanning and Processing locations, Persons involved etc.	Identification of bottlenecks, calculating computing efficiency and report generation
Structural Metadata	Size of each page, Blank Pages, Page Context - Beginning of Chapter, End of Chapter, Index, Preface, Table of contents etc.	EasyNavigation, better search and retrieval

Table 1: Category and type of information in the metadata

3.2 Groups in Process

The process involves different stages of procurement, metadata entry and handling and quality assurance. To execute these stages the process consists of different groups.

3.2.1 Procurement Group

Procurement group identifies and procures rare monumental works and approves, based on the content, the books that need to be digitized. For such a committee the understanding of usage statistics of the already online books is quite essential. The procurement group also takes care of copyright and intellectual property issues. The authors are provided due respect for the book, by placing the acceptance for web-enablement form on the first page of the book. This has encouraged many authors to accept digitization process and offer the books for web-enablement.

3.2.2 Metadata Group

This group consists of librarians and technicians and is responsible for the entry and validation of the metadata. Since books are scanned from multiple locations and libraries through out India, we need librarians who can understand different languages and have a diverse knowledge on various subjects. Usually the metadata is verified and corrections are made by remotely distributed librarians who can log into the system and make necessary corrections over the web.

1. **Manual Verification:** Manual verification of metadata takes place at the RMSC end. Once the book is on the DLI servers, expert librarians can access the book online and verify the metadata. For books belonging to multi Indian languages, the entry of titles should be in a standard searchable format free from the fonts displayed [6]. In DLI we follow the ITRANS format with support for open and true type fonts.

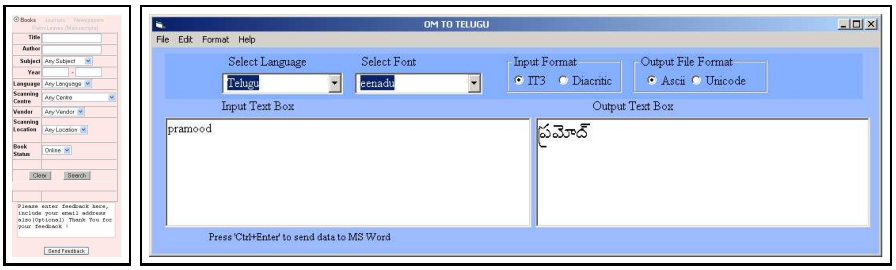
2. **Automated Check:** Regular metadata fields like Title, Author, keywords, and subject are filled in for most of the books. In a large portion of the books they are misspelt, incorrect or do not exist. Technicians from the metadata team run automated categorizer to categorize the books into their respective categories/subjects, which are then clustered and corrected by the expert librarians.

3.2.3 Quality Assurance Group

This group verifies the digitized content and approves it for uploading and hosting on the web. They perform the check for duplicates, damaged pages, missing pages, file formats and also on other parameters to ensure that the quality standards are met. Administrative issues regarding the decision making of the undefined errors found in the digitized books and content is also made by this group. The group also ensures the process is carried out in the defined manner and performs process audits for applying the improvement strategies.

4 Workflow

The books that are to be digitized are procured and handed over to the librarian. The digitization of a book starts with the librarian entering the regular metadata for the books that need to be scanned. If a book has already been scanned it should not be re-scanned again at any other location. Hence, the metadata is first uploaded onto the DLI portal hosted at a local Mega Center for checking of possible duplicates from elsewhere. The portal consists of a central repository for the DLI where all metadata from different mega centers is aggregated and stored. The book to be scanned is checked against all the exisitng books in the repository for possible duplication. If the book already exists, it is not cleared for further action. If no match to the book is found, the book is allotted to a vendor/contractor for scanning and processing the book. Also, the books are checked for copyright conflicts and only those which do not have any, are cleared.
As shown in the Figure 2(a) the librarian enters the metadata information into the DLI server. The server then checks if hte book is already processed. if not the book is cleared for digitisation. In case of Indian language books, the OMTrans transliteration scheme is used to encode the Indian language names, titles etc to Roman representation. The tool used for this purpose is shown in Figure 2(b).



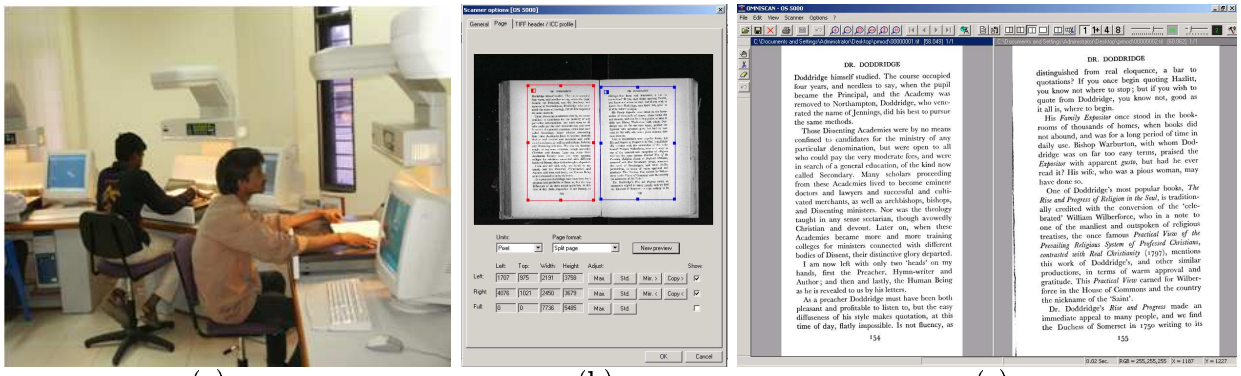
(a) (b)

Figure 2: Meta Data Entry Form is filled before the scanning (a) using OMTrans Transliteration(b)

4.1 Scanning

Scanning of a book is at the heart of the entire DLI project. It requires the maximum resources both in equipment and man power. This process needs quick,high quality and very reliable scanners for scanning (or photographing) the books. It also needs well trained personnell to operate these scanners efficiently and to achieve maximum output. The scanners used at RMSC-Hyderabad are made by Minolta and Zeutschel. The Minolta scanners scans a vertical strip of the spread of the book and many such strips are accumulated to form the final image, while the Zeutschel scanners take a high resolution photograph of the spread giving us the scanned image of the spread.

These scanners come with associated drivers. The software generally used to scan the pages is ABBYY Fine Reader or OmniScan. Using this software the scanning operator sets the various parameters for the scanning process such as the "dpi". Dpi or Dots per Inch specifies how many pixels (or dots) in the scanned image will constitute one inch of length (or width) in the actual page. A high dpi implies high resolution and thus high quality. However, high dpi also means the images need more storage space. Taking care of both quality and storage constraints, the DLI has specified a minimum scanning resolution of 600dpi.



(a) (b) (c)

Figure 3: Various Stages in the scanning Process: (a) Scanning in Progress; (b) Configuration-settings of the scanner. (c)Screen shot after scanning a spread.

After setting the dpi, the scanning operator sets the approximate page boundaries. This is needed so that the software can automatically separate the left and right pages from a single image of the entire spread. To set this, the operator first chooses an arbitrary spread which

represents the size of textual content in the book. He then scans this spread and specifies the approximate bounding rectangle of the left and right pages. He verifies and corrects this bounidng rectangle by testing it on a few other arbitrary spreads as shown in Figure 3(b).
The operator then proceeds to scan the entire book using these settings. Figure 3(c) gives a snap shot of scanning a spread using OmniScan. The scanned page is saved as a TIFF Image and compressed using the CCITT 4 Fax Compression scheme. Each scanned page is named as an 8 digit number like 00000001.tif and so on. These scanned pages are stored in the OTIFF folder of the book. Optionally the pages are separated as even and odd numbered pages for convenience during the image processing stage, and they are stored in separate folders. However, when the book is finally being submitted to DLI, such folders should be removed.
After the scanning of the book is completed, the operator does a superficial check whether all pages have been scanned properly. Now the book is sent for Image Processing and Optical Character Recognition(OCR).

4.2 Image Processing

During this stage the images of each scanned page are cleaned, cropped and processed so that the images are human readable and also can be well processed by the OCR. Firstly, the image has to be cropped. Cropping is the operation of extracting only the relevant portion of an image and removing the rest. In the scanned image there will be dark areas on the edges of the page and sometimes the fingers of the scanning operator are also photographed. These are to be removed from the textual content of the image. To do this the operator chooses a full page and fits a rectangle to the textual portion. All other pages run through this rectangle and anything lying outside this rectangle is deleted.
The operator then runs various Image Processing operations [3], using the Scanfix software. Before the operator runs every image through Scanfix he sets the following parameters:

Noise: in many books, especially old ones, we can see that the pages become dark and a number of dark spots appear in the page. When scanned this forms the noise in the image. This is removed by setting a noise removal of 6% de-speck value. If we increase the de-speck value, it might remove some of the text like the period(or full-stop) considering it to be a noise spot. In case there is still some noise after the Scanfix is run, the operator removes them manually by using the erase tool.

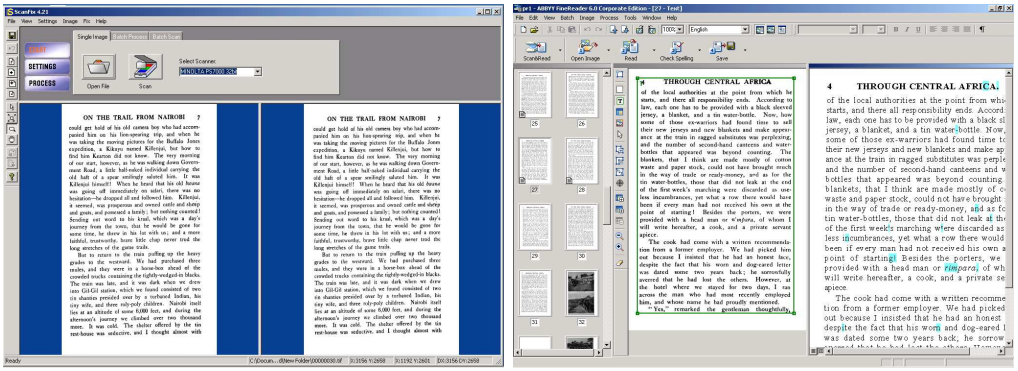
De-Skew: Skew is the rotation present in an image. When scanning tightly-bound or loosely-bound books, the pages are not aligned perfectly. So an amount of skew is found in the picture. The de-skew operation is to correct such skew in the image. To do this the operator selects an approximate bounding rectangle around the content of the page and if any page has content outside this rectangle, the content is rotated and resized to fit within this rectangle.

Intelligent Crop: Provides a uniform margin around the content of the page. DLI specifies that a minimum of 300 pixel margin has to be there for all pages on all sides of the page.

Smoothness and Completion: In case of heavy noise in the images some of the characters are either cut or merged. To correct these, various morphological operations such as Dilation, Erosion and sandfill are used. The operator sets the parameters for each of these operations.

Resize: To maintain uniformity in all the images of a book, all pages should be of the same size or Height x Width. The operator chooses an arbitrary page and sets its dimensions as the standard for all pages and all pages will be resized to this standard size.

Once these parameters are set, the Scanfix software applies these parameters to every page in the book as shown in Figure 4(a). In case the Scanfix software couldnt process any page satisfactorily, the page is sent for rescanning and the rescanned image is processed. The resulting images are the processed TIFF images and are stored in the PTIFF folder of the Book Folder.



(a) (b)

Figure 4: Screen shots from the softwares used for image processing and character recognition. (a) Image Processing using Scanfix (b) OCR using ABBYY Fine Reader

4.3 OCR

OCR or Optical Character Recognition is the process of recognising characters from scanned document images. In this process, an image with textual content is read and the characters present in the text are outputt by the OCR software. The OCR software being used at IIIT-Hyderabad is the one provided with ABBYY Fine Reader. The OCR takes as input the images in the PTIFF folder of the book and generates text which is stored in three formats : HTML, RTF and TXT. For every page that is OCRed we have 3 files generated in each of these formats and are respectively stored in the folders named HTML, RTF and TXT. In Figure?(b) we show the snapshot of ABBYY Fine Reader when it processed an image and output the text int he image.
The ABBYY Fine Reader uses special recognition technology based on the principles of Integral Purposeful Adaptive (IPA) perception. Some of its features that make it very convenient to use are:

- Omnifont: recognizes texts in practically any font
- High recognition accuracy
- Low sensitivity to print defects
- Recognition of poor print quality documents
- 177 recognition languages (doesnt include Indian and Chinese)
- High speed recognition

- Can run in batch mode
- Layout Analyser: text, tables and images displayed in their original location
- Saving of non-rectangular images, multi-column text flows and list

After OCR is performed the book with all its contents is sent for the Final Quality Checking and submitted to the Mega Centre for being hosted on the web.

4.4 Quality Check

When a book is submitted to an RMSC for Quality Checking, it is first checked for duplication in the existing collection of books. such duplicates are displayed so that action can be taken by the Quality Assurance team and such dupliactes must be removed from the server to avoid wastage of storage space. The tools developed at IIIT-Hyderabad can intelligently search for duplicates by doing partial string matching on the book title, author, publisher and other significant fields and returns a list of books that are duplicated, as shown in Figure 6(a)
After checking for duplication, the book is checked for its quality. Quality Check is one of the very imporatant stages in the DLI workflow. Its essential to maintain uniformity in the entire digital content and to ensure good quality in the deliverables of the project. DLI has speicified a set of very strict guidelines on quality parameters. Every vendor, contractor or centre should adhere strictly to these norms and ensure that high quality material is generated form their respective centres.
The contents of a book are as shown in Figure 5.

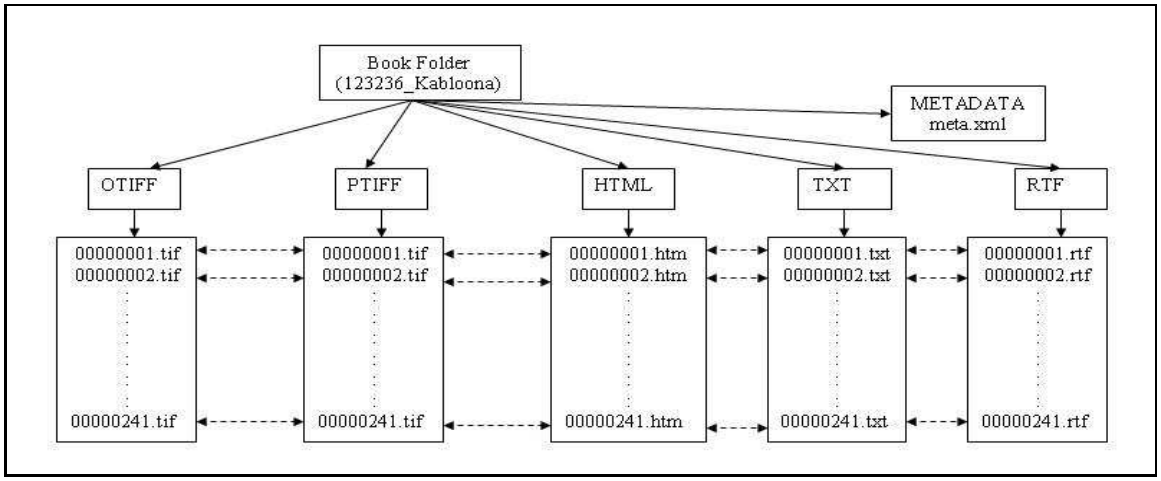


Figure 5: Contents of a Digital Book is organised as a tree in the Hard Drive. Content is stored as images, text and html in separate directories with appropriate file names.

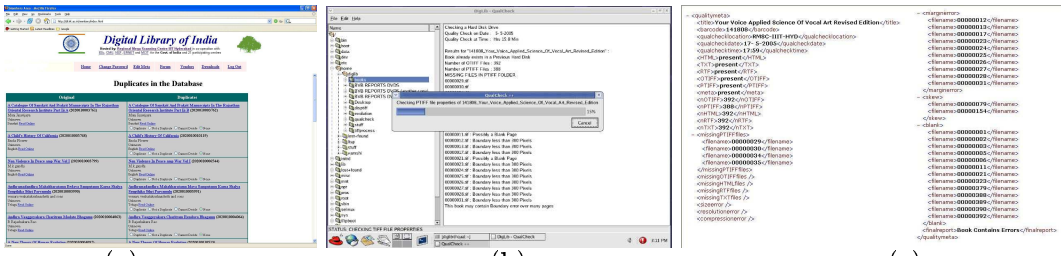
The items in the book are the Metadata XML file and the 5 folders namely OTIFF, PTIFF, HTML, RTF, TXT. Every page in the book, including blank and missing pages, has a file associated with it in each of the folders. The corresponding files in different folders are given the same name except for the extension. Apart from these specified items, there should be no stray folders and no files should be out of place. However, incase an OCR is not available for

the language of the document, the three folders HTML, RTF and TXT will not be present. During Quality Check the presence of each of these items is checked and the missing/stray items reported accordingly.
The processed scanned image of each page, namely the files in the PTIFF folder are considered the most valuable output in the entire scanning process. The quality parameters which the PTIFF files should meet are as shown in the table:

Dimensions	Same size (height x width)
dpi	600 or above
Compression algorithm	CCITT 4 Facsimile Compression
Margin	300 pixels on all sides of the page
Skew	$< 2°$
Blank Pages	Should be identified and annotated

Table 2: Major Quality Parameters for Processed TIFF Images

During the Quality Check, each of the PTIFF folder files is checked for each of these parameters and the discrepancies reported accordingly for correction. The final report is stored in an XML file called qualmeta.xml. The QualCheck tool which was developed by IIIT-Hyderabad automatically checks for presence of errors in the submitted books. The tool recursively searches the HDD/DVD for books and generates the XML reports of the quality parameters, as shown in Figure 6(b). A sample report is presented in Figure 6(c).



(a) (b) (c)

Figure 6: Screen shots from the inhouse developed tools for Quality Assurance: (a) Duplicates Detection (b) Quality Ckeck of the Content (c) Quality Check Results as XML

4.5 Web Enablement

A book cleared by the Quality Check stage is submitted to a Mega Centre that hosts the books on a webserver. The book is checked for duplication in the server. An operator performs the post-scanning metadata process. This gives us the structural metadata as described in the section above. A copy of the book is made and stored as a backup in case of a hard disk crash. The Mega Centre puts these books on the sever and also duplicates it for the local servers of other Mega Centres. The details of web hosting are presented in the next section.
The entire process flow can be represented as shown in Figure 7.

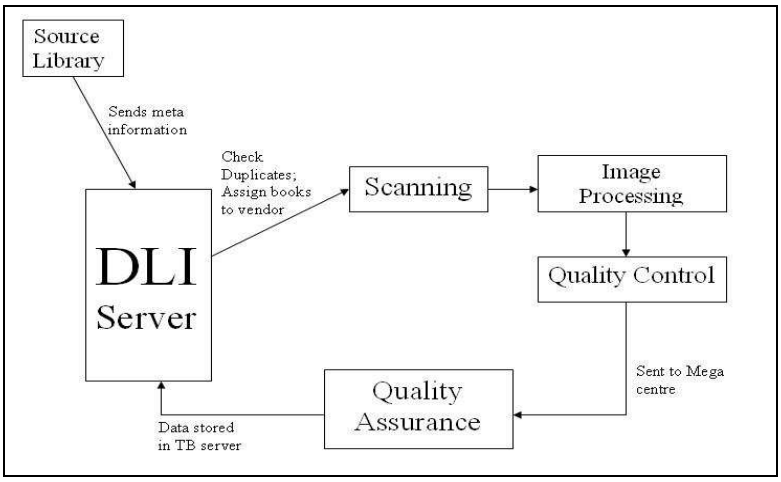


Figure 7: DLI Workflow

5 Architecture of DLI Server

In this section we describe the architecture that supports the process and the model discussed in the earlier sections. Many issues of coordination have been resolved by modularizing the tasks with the help of technologies like XML and Databases and discrediting and distributing the efforts in the project as much as possible. Web services have proved increasingly useful in solving coordination problems and are gaining popularity in the enterprise world.

5.1 Software Architecture of the Mega Centre Digital Library

At RMSC Hyderabad we had experimented a few different archtitectures before we finalized upon the one shown in the figure. The current architecture is motivated by factors like scalability, ease of maintenance, dependability and economy.

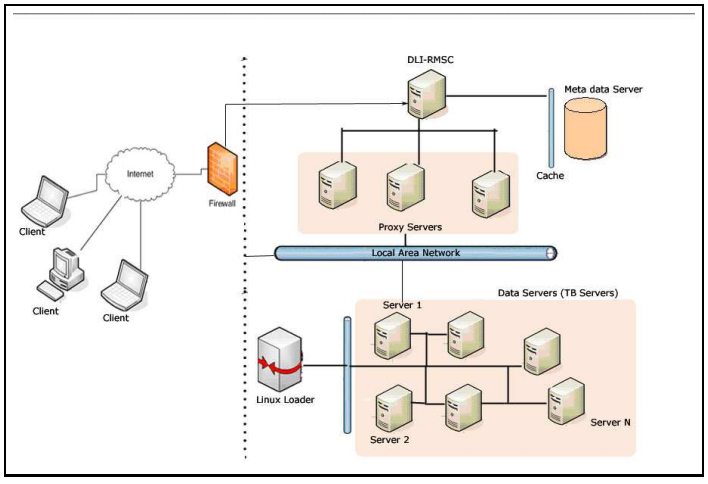


Figure 8: a. Architecture of the DLI hosted at a Mega Centre b. Decentralized SOA based Architecture of the Central DLI

The digital objects are preserved on Terabyte servers which are clustered as a data farm. Each server in the data cluster hosts all the digital objects preserved on it, through an Apache web-server. The cluster is powered by a distribution of Linux that is enhanced to support diskless network booting. This option of diskless network booting helps us boot a server without having to devote any space for storing the system specific and operating system files. This set up is economical and and also easy to manage, in a way that we can add or replace data nodes in the cluster instantaneously without the hazzle of operating system installations.
Data can be copied onto a server in the cluster over the network or through USB interface. The servers implement a hardware RAID to contain disk failures which adds to the reliability of the system. Also a redundant copy of the complete data is present on external storage media, for data restoration in the event of irrecoverable crashes.
The 'metadata server' is a repository of the complete metadata which is in XML. XML has been chosen for its important role in interoperability. Metadata is passed on constantly between contractors and the RMSC, and it also acts as an identifier of the book that is to be scanned. Using XML as the format, modularises the work by decoupling RMSC and contractors and also ensures smooth interoperability. Wrappers present on the metadata server automatically populate the database from the xml metadata. When metadata is uploaded onto the server, it is first checked for duplicates on the existing database. If no duplicates are present, the book is permitted for scanning.
Once the scanning is completed the book content is returned by the vendors. The books are then uploaded onto the data farm and the metadata for each book in the database is edited to contain a pointer to the location of the book in the data cluster. The portal has a front end using which a user can login and query on the metadata to retrieve books he wishes to read online. A caching mechanism deployed on the metadata server helps us cache similar queries posed to the database and return the results promptly. When a user requests to view the complete book content, the location of the book in the data cluster is gathered from the database and content is retrieved over http requeusts, from the particular server in the cluster and is broadcast to the user. In this way the main portal only acts as a Proxy server between the user and the book server that contains the requested book.
The features of this architecture are such that

- The Architecture balances resource usage within the community
- It has high data availability
- The data is accessible even if creator disappears from the system
- Is easily updateable such that the stored data can be modified during the system lifetime
- It supports a powerful query language like XPath / XQuery

6 Summary

The Regional Mega Scanning Center at Hyderabad was established as per the direction of the Ministry of Communication and Information Technology and has been operational for over an year. The Center is active in content creation for the digital library, building test-bed for research and carrying out research on important technologies associated with digital libraries. Center operates at IIIT and multiple subcenters with about 50 high speed scanners. Content at Salarjung Museum, Osmania Univ., Telugu Univ, State Central Library and City Central

Library etc. are getting digitized as part of this project. More than 15 Million Pages are hosted on the DLI server. RMSC has established a smooth workflow procedure and we have generated a large amount of digitized content. This content is being hosted on reliable webservers which currently host about one lakh books online.

Acknowledgements

We acknowledge Prof. Raj Reddy and Prof. N. Balakrishnan for their vision of the DLI project and for having laid down procedures for the establishing and running the DLI. We thank Prof. Rajeev Sangal and Dr. C. V. Jawahar who have constantly guided and supported the RMSC at IIIT Hyderabad. We profusely thank Mr. Kiran Kumar and the vendors at RMSC Hyderabad namely PAR Informatics and Trinaina Informatics Limited for their cooperation and hard work, which has made RMSC Hyderabad the biggest content development centre for DLI. We also thank the many people who were deeply involved at various stages, the librarians, the data entry operators, the scanning operators, image processing personnell, server maintenance executives and administrative staff at RMSC Hyderabad, for their involved and dedicated efforts towards the DLI project.

References

[1] Ingo FromHolz, Predrag Knezevic, et al: *Supporting Information Access in Next Generation Digital Library Architectures*, Proc. Sixth Thematic Workshop of the EU Netwirk of Excellence DELOS(2004)

[2] M. E. Lesk: *Understanding Digital Libraries* Morgan Kaufmann, 2004.

[3] Gonzalez, R. C., Woods, R. E., *Digital Image Processing* Prentice-Hall, 2002

[4] Raj Reddy: *The Universal Library: Intelligent Agents and Information on Demand* ADL 1995: 27-34

[5] http://dli.iiit.ac.in

[6] http://www-2.cs.cmu.edu/ madhavi/Om/