Digitizing a Million Books: Challenges for Document Analysis

K. Pramod Sankar[1], Vamshi Ambati[2], Lakshmi Pratha[1], and C.V. Jawahar[1]

[1] Regional Mega Scanning Centre,
International Institute of Information Technology,
Hyderabad, India
jawahar@iiit.ac.in
[2] Institute for Software Research International,
Carnegie Mellon University, USA

Abstract. This paper describes the challenges for document image analysis community for building large digital libraries with diverse document categories. The challenges are identified from the experience of the on-going activities toward digitizing and archiving one million books. Smooth workflow has been established for archiving large quantity of books, with the help of efficient image processing algorithms. However, much more research is needed to address the challenges arising out of the diversity of the content in digital libraries.

1 Introduction

With the increased availability of economical digital storage media, high speed scanners and high-bandwidth networks, Digital Libraries have received a boost in the last few years. The dream of digitizing the vast knowledge of mankind, and making it available online has now become a realisable goal. As a first step towards this goal, the Digital Library of India (DLI) [1] and the Universal Digital Library(UDL) [2] projects aim at digitizing *one million* books and making them available on the web. One million books is less than one-percent of all the world's published books and represents only a useful fraction of those available.

However the digitization of a million books is in itself a herculean task. If on an average, a book contains 400 pages, then the project will create 400 million digital documents making it the single largest collection available on the web, with as many documents as a tenth of the number of web pages over the entire Internet. If it takes only one second to digitize a single page, it would require about a hundred years of time and 150,000 GB of storage space for the project. Digitizing such massive quantity of data and making it available on the web, for free and non-stop access to anybody-anywhere is clearly a stupendous goal.

The vision of digitizing books was originally conceived at Carnegie Mellon University [3]. A 100 book and a 1000 book pilot projects were successfully completed [4], which paved way for undertaking the goal of 1 million books by 2008, called the Million Book Project (MBP). The project is presently pursued by multiple institutions worldwide in the United States of America, India, China

H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, pp. 425–436, 2006.

and many other countries. The Digital Library of India (DLI) initiative is the Indian part of the UDL and MBP. As of now, the Regional Mega Scanning Center at IIIT-Hyderabad, and associated digitization centres have contributed a major portion of the total content generated within India.

The entire process of digitizing the book consists of various stages such as procuring, scanning, image processing, quality checking and web hosting. Some of the issues involved in creating digital libraries are given in Lesk [5]. The digitization of a million books in a realistic time frame requires the state-of-the-art scanners, and high speed algorithms and software to process the scanned images. An efficient system needs to be built, that pipelines these processes and ensures smooth running of the project, enabling timely delivery of digital content. The system should be robust to various logistical, technical and practical difficulties in handling the work at this large scale. We have succeeded in developing such a system, where a large number of books are digitized each day. The process runs in a highly distributed and layered environment, yet the data flows freely from one stage/centre to another. The statistics from this approach have proved that the envisaged dream could be realized very soon.

The UDL/MBP is the first digitization project aiming at such large quantity of digitization as a million books. Recently, many new projects are being undertaken across the globe, which also aim at digitization of libraries. These include projects like American Memory, Project Gutenberg, Gallicia run by Frances national library, the University of Pennsylvanias Online Books, California Digital library, etc. Project Gutenberg, founded in 1971 by Michael Hart, and built and maintained by hundreds of volunteers, is the longest-running project. As of date it has reached the figure of 16,000 books. Organizations like NSF and DARPA, sponsor a large number of digitization projects in the USA. These initiatives are generally specific to a particular domain. Many projects have recently begun in Europe where the emphasis is on European information resources combining multicultural and multilingual heritage in Europe. Both digitized and born digital material are covered by this initiative. Companies such as Google, Yahoo and Amazon have also undertaken digitization of books on a large scale.

Most of these digitization efforts are either relatively small or too often associated with restricted access. Google Print, for example currently restricts access outside the US even to titles that are in public domain outside the US. Within the US they have a restricted access and some roadblocks to printing and saving images even to public domain titles. Gallica's interface might not be the easiest to use for non-French speakers. The University of Pennsylvania online books project also has to work out on quantitative and quality control issues with respect to its scans, metadata, and interface. The efforts of commercial ventures are associated with a proposed business model to allow for shopping of books based on snippets of content and to reward publishers and authors of copyrighted material.

DLI, however, is a non-commercial project aiming at digitizing non-copyrighted books, mostly of Indian origin. Copyrighted books are digitized and made available online, only with written permission from the author and the publisher. Apart

from printed books, a large number of palm leaf manuscripts, which are a part of the Indian heritage, are also being photographed and digitized for the preservation of such delicate storehouses of ancient knowledge.

An insight into the challenges faced by the document image analysis (DIA) community in building digital libraries is discussed in [6] [7]. In this paper, the various challenges realized from the actual implementation of the MBP are discussed and an overview of the procedures employed to work towards the goal is presented. The rest of the paper is organized as follows. Section 2 describes the various issues that affect the digitization process at this large scale. Section 3 gives an overview of the process workflow which is a semi-automatic document analysis system. Section 4 charts the progress and performance of the system at work in the RMSC at IIIT-Hyderabad. Section 5 observes the challenges that we realized while working with the project and we conclude in section 6.

2 System-Level Issues

When the Million Book Project was initiated, it was the first of its kind ever conceived. The challenges faced by the project were many, and as the project progressed, newer challenges arose. The major aspect of the project is its magnitude. Digitization of a million books, firstly, requires being able to procure the given number of books. The type of books could vary in a variety of ways such as the size of the book, quality of paper, clarity of print, font face used, type of binding etc. Old books which are delicate and have deteriorated over time need extra care in handling apart from requiring special routines to process them.

Transferring the procured books to a digitization center calls for heavy logistical inputs. To reduce this, the digitization centres are generally setup in close proximity to the libraries. Such centres should be co-ordinated and managed by Regional Mega Scanning Centres (RMSCs), where the digital content generated by each center is hosted on servers for public access. The storage and transfer of the data generated by each center is a major task. Each center generates tens of GBs of data every day. Since network transfer of such large quantities of data is neither feasible nor economical, it has to be physically moved in HDDs or DVDs. Pipelining the data within the various phases of the digitization process is also a serious tactical issue.

Due to the fact that different libraries have a copy of each of many books, duplicate digitization occurs, which results in wastage of effort and thus valuable resources. To avoid this, there has to be proper synchronization between the various digitization centres and RMSCs. Moreover, the representation of Indian languages content is not standardized. Different organizations use different formats, such as ISCII, ITRANS, OmTrans, Unicode, etc. to code the Indian language documents. This lack of a standard complicates duplicate checking and elimination.

Digital Libraries are also a means of preservation of content for the use of future generations. Hence maintaining good quality in the digitization process is of utmost importance. Also to ensure that all centres adhere to a common set of guidelines, enabling and maintaining a standard [8] for the entire project is essential. Baird [9] gives some guidelines regarding many of the quality control aspects for digital libraries.

Maintaining large amount of digital data on the web, for any time access from anywhere, is a huge challenge. Accurate search and quick retrieval of the digital content against user queries is a major research problem.

Considering these issues, a process workflow was designed and is explained in the next section. Apart from handling the above problems, the system was also robust to many practical situations which threw up further challenges.

3 Semi-automatic Digitization Process

Owing to the magnitude of the project, the digitization process is distributed into different logical steps and over different *wings* within a given digitization center. Since, a distributed environment is apt for the project, we have identified different phases of work that can be distributed over different locations. We designed a process flow that was aimed at achieving a highly automated set up and to create the notion of a distributed environment, which is flexible, yet cohesive. The different stages of the process pipeline are depicted in Figure 1. Each of the component of the workflow is explained below.

The digitization process begins with the procurement of books. A librarian creates the metadata for the books, which is checked for possible duplication against previously digitized books. We have built technology and solution [10] to avoid duplication within our center with minimal network resources.

Scanning: The digitization process starts with the scanning of books. High speed scanners are used to convert books to the corresponding page images. Overhead

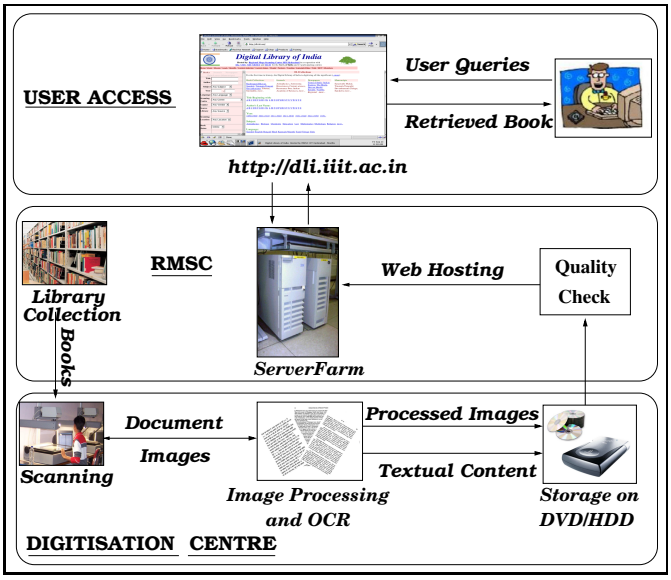

Fig. 1. Overview of the Digital Library Process

scanners scan or photograph the spread of a book from above and can typically scan an A4 sized page at 600dpi in about 2.5 seconds. Theoretically, such a scanner could scan about 10000 pages in a 8 hour shift day. However, due to the fact that an operator needs to turn the pages of the book and give the command for a fresh scan, the obtained throughput is about half this number. In the special case of digitizing palm leaves, high resolution digital colour cameras are used to photograph them.

Image Processing: The raw images generated from the scanning need to be processed for improving the quality. The scanned image needs to be cropped to remove the background. The textual content of a page generally contains a number of artifacts in the form of dots and blotches on the page due to aging of the paper or moth bite, tear or cut in the page, and eroded or incomplete characters. Various image processing operations such as cropping, de-skewing, de-noising, smoothing etc. [11] provided by the ScanFix software, are performed to rid the images of such blemishes. Using this software the operator identifies the most appropriate steps to remove the artifacts as much as possible. He also sets various parameters for the image processing procedures.

The processed images are stored in the TIFF file format using the CCITT 4 Fax compression algorithm. This scheme was found to be very efficient. An image of 4300×2900 resolution, containing only a few words is of 1KB file size, and a document of the same resolution with full text occupies 75KB while the same would need about 150KB in PNG format. A page with images stored in binary format would be of 120KB in TIFF format, compared with about 230KB in PNG format.

We have also experimented with the image processing of palm leaf document images. The software used for books does not support the required functionality for processing palm leaf images. Unlike the white background for book pages, palm leaves have a dark brown background which is non-uniform. Owing to the brittleness of the medium, the documents are prone to tear and damage and the text is found with very heavy noise. No commercially available software is able to handle such extreme conditions and special software is being developed for this purpose. A sample palm leaf and its processed version are shown in Figure 2.

Recognition and Reconstruction: After cropping and cleaning a page image, an OCR is used to extract the text from it. A commercial OCR available with ABBYY FineReader is used for this purpose. The text output by the OCR is stored in RTF, TXT and HTML formats, in separate folders. The extracted text is used to index the pages and books, to enable search and retrieval for the users.

Quality Checking: To ensure that proper quality is maintained by the digitization center, the digitally converted book is checked for quality at the RMSC [12]. A set of fixed standards were adopted as given in Table 1. The submitted books are checked for quality in each of these aspects and books with more than 80% of the contents containing errors are re-scanned and/or re-processed.

An automatic tool, *QualCheck*, was developed, that searches recursively for books in a given HDD/DVD and automatically checks for all the required con-

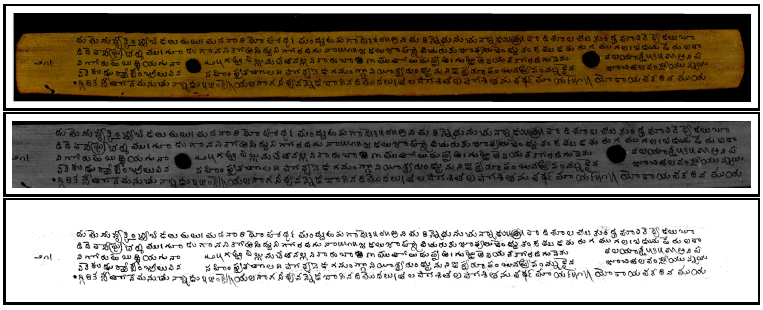

Fig. 2. Palm Leaf images a)unprocessed image b), c) processed images

Table 1. Major quality parameters for processed TIFF images

Parameter	Specification
Dimensions	Same Rows×Columns
dpi	600 or above
Compression	CCITT 4 Facsimile
Margin	300 pixels on all sides
Skew	$< 2^\circ$
Blank Pages	identify and annotate

tents. It then checks for each of the parameters defined for the images, and generates an XML report for each of the errors. A snapshot of the tool is shown in Figure 4 (a). Typically observed errors are missing files from one or more of the folders, non-uniform margins in the processed images and use of a different compression algorithm than specified.

Web-Hosting: The digital book is hosted at the RMSC on Terabyte servers which are clustered as a data farm. An operator performs the post-scanning metadata process, creating structural metadata [10], which is used to easily navigate through the book. A copy of the book is made and stored as a backup in case of any hard disk crash. These books are hosted on the sever and duplicated for the local servers of other RMSCs.

4 Performance of the System

The process flow described in Section 3 was implemented at the RMSC-Hyderabad and the results were very satisfactory. The center was able to achieve a throughput of 140000 pages (about 500 books) each day. Observing the efficiency of the workflow at this center, many other centres have adopted the same pipeline to establish the process at their locations. The following are the performances of each phase of the digitization process.

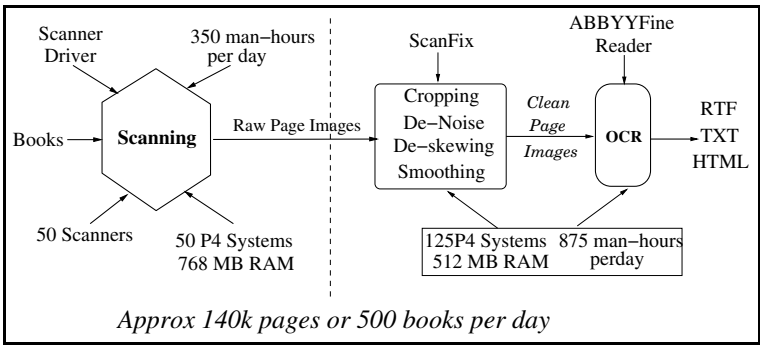

Fig. 3. Scanning Process, Image Processing and OCR schematic diagram with inputs, output and throughput

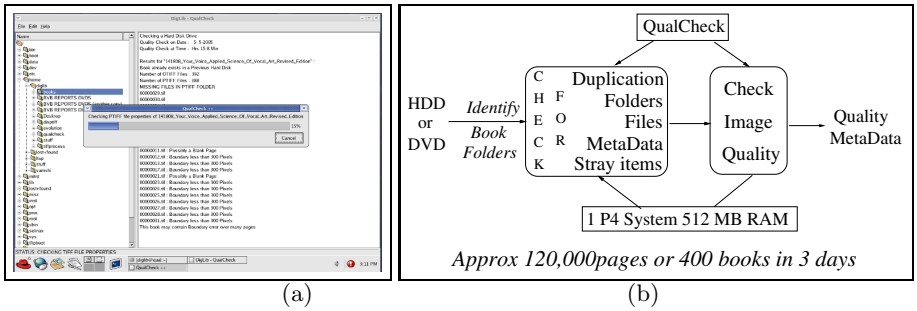

Fig. 4. (a) Snapshot of QualCheck tool (b) quality check performance

Scanning: Figure 3 summarizes the scanning and the image processing stages along with the quantity of inputs and outputs per day. The left side of the dotted line represents the scanning phase. At RMSC-Hyderabad, about 50 scanners are operated in an 8 hour shift per day. The peak throughput obtained was about 140000 pages per day, or 500 books. At this rate, it would take 10 years for a single center to scan a million books. Thus, to increase the overall throughput, the setting up of many such centres, at various locations, is inevitable.

Image processing: The Image processing and OCR stages along with the metrics are depicted on the right side of the dotted line in Figure 3. It was observed that a single desktop PC, can process about 2000 pages per day, performing both image editing and OCR. To match the throughput of the scanning phase, the number of systems allotted for the image processing phase is $2\frac{1}{2}$ times the number of scanners. At RMSC-Hyderabad, the image processing facility has an output of 150000 pages per day, with 125 machines being operated over a 8-hour shift per day.

Quality Check: The Quality check process is depicted in Figure 4 (b). The input to the process is a DVD or HDD containing books submitted. The QualCheck tool needs minimum user input. Once submitted, a HDD containing 400 books

Table 2. Range of digitized content created

Aspect	Diversity
Total Books	90,336
Pages	27,856,099
Publication years	1852-present (for Books)
Medium of books	Paper, Palm leaves, Cloth
Types of books	Printed, Handwritten, Engraved
Languages	Arabic, English, French, German, Greek, Italian, Norwegian, Persian, Spanish, etc.
Indian Languages	Bengali, Hindi, Kannada, Marathi, Sanskrit, Tamil, Telugu, Urdu, etc.
Subjects	Art, Architecture, Autobiography, Astronomy, Commerce, Religious, Economics, Science, Engineering, Geography , Law, Health, History, etc
Print Quality	Press, Offset, Newsprint, Journal, Electronic printing
Sources of Books	State Libraries, Universities, Museums, Religious Institutions

would be checked in close to 3 days or 72 hours. However, the process slows when the books are submitted in a DVD because of the slow data transfer from the DVD hardware. With three systems performing the quality check over HDDs, the throughput of this stage is matched with the previous stages.

4.1 Present Status

Using the procedures stated in this paper, we were able to achieve high quality output and high throughput from the digitization process. In just over an year, RMSC-Hyderabad digitized more than 100,000 books which contain about 25 Million pages. The books are currently online at [1] and are available for free access to anyone, anywhere in the world. These books cover a range of subjects, and languages. The diversity of the digitized content is showcased in Table 2.

More than one third of these books are of Indian language content, spanning 8 Indian languages. Since no commercial OCR is available for Indian language character recognition, the textual content is not available for these books. This severely handicaps searching, which is restricted only to the title and keyword level search. Special techniques need to be developed to search within the books.

Most of the books that are available online from the DLI, date back to before 1920. Many of these books are out of print and could be the last available copies of the same. As part of the initiative to prevent valuable information from being lost, we digitized about 4000 books from the Salar Jung Museum in Hyderabad with more in the process. We are also undertaking further procurement of books from rare collections.

Palm leaves: Digitization of palm leaves is a first-of-its-kind activity undertaken by us. Palm leaves were the storage medium for ancient literature, philosophy and science containing valuable knowledge. The preservation of these manuscripts is

one of the more respectable achievements of the project in the context of cultural preservation, apart from other benefits discussed earlier.

5 Challenges Ahead

Thus far, all the books were scanned in binary mode only. This is valid because most of the books being scanned do not have any colour content. Digitizing books in colour would require more sophisticated scanners and better image processing and recognition algorithms and systems.

Trained Manpower: One of the bottlenecks of the project was the lack of trained personnel to operate the scanners and perform the image processing tasks. Since the equipment are costly, much care has to be taken to properly utilize them, and ensure high output without causing any damage. To this end, the operators were put through a thorough training program. The resources required for training could be reduced by using software that is more user friendly.

Duplication: Duplication of work was a major problem that reduced the total output. Since different libraries have a copy of each of the many books, the same books were digitized at different centers. To avoid such duplication of effort and wastage of resources, different meta data files were specified and procedures were laid out [10]. This addresses part of the duplication problem.

Quality of Metadata: Metadata management is an important aspect of the digital library. Frommholz et al.[13] signify this aspect from an information access perspective. A major portion of the sources of books in the project have metadata only in non-digital formats and these have to be entered manually. For the entry of metadata of a book, we largely rely on the librarians for the accuracy and credibility. Sometimes, librarians might not be well advised about the hierarchy and ontology of book classification. We ascertain the fields of metadata by referring to the catalog of OCLC and also by machine aided manual correction of metadata.

Indian Language Content: Most of the established softwares and routines are tuned to handle documents set in the Roman alphabet. The Indian languages, however, are very different, having a large character set and such features as *matras, samyuktakshars, shirorekha* etc. This results in conjunct and compound characters with complex shape variations. In addition, Indian language processing is considerably complex compared to English. These factors complicate optical character recognition, search and indexing. Kompalli et al.[14] give an overview of the challenges in OCR for one of the Indian language script called *Devanagari*. Thus, better software needs to be written to handle Indian languages.

The lack of a standard format for representing Indian language content is another handicap. A standard format needs to be developed and appropriate converters must be made available for converting from one format to the other. Software has to be developed using this standard. The non-standard representation also hampers the web enabling of digital content from scanned books. Web browsers and book viewers should be made "Indian language enabled".

Robust OCR: The OCR software being used gives an accuracy of about 90% - 95%. However, this accuracy is not good enough for a powerful search engine to be built for the digital library. Further improvements in OCR technology are being awaited. In case of noisy images, the performance of the OCR degrades rapidly. Some of the approaches to deal with noisy images are given in [15]. Moreover, OCRs for the Indian languages are still in the research phase, and satisfactory systems need to be built for OCRing Indian language content.

Search in Presence of Errors: Due to the inherent limitations in the OCR design and performance, the accuracy obtained is limited. The commercial OCR being used produces about 5% of errors. If a page contains about 2000 characters, then in a book of 300 pages we could find more than 30000 erroneous characters, which could mean that many erroneous words in a single book. The situation is worse in case of Indian language OCR where the errors are at the component level of each *akshara*. The errors in the component recognition cascade into erroneous letters, and then words. Inspite of these errors, we need to be able to search the content. To achieve this, search in presence of errors needs to be addressed, unlike the exact search that is used so far.

Document Image Compression and Delivery: As was showcased earlier, the storage requirements of the digital content generated from the project is enormous. The storage of such large amounts of data, reliably, is a huge task. With the OCR performance not satisfactory as described above, the preferred delivery of documents will remain to be images for the immediate future. Thus the file size reflects the network bandwidth required for the transfer of page images. One way to reduce the storage requirements is by developing better compression techniques that are tuned for document images. Better bandwidth availability would also improve the utilization of the Digital Libraries by the common man.

Historic Documents: Digitizing palm leaf documents has a large set of challenges associated with it. They need to be handled with enormous care as they are delicate and irreplaceable. As of now, the palm leaves are being photographed one at a time, but better methods have to be invented to improve the rate and quality of digitization. Better scanners have to be designed for this purpose, which can take high definition colour images of the palm leaves while causing no damage to the bundle. The processing of the palm leaf images needs special features and the image processing algorithms must be well tuned to handle heavy noise, tear, cut and background features. Currently, image processing of palm leaf images is pending the appropriate software development.

Human-Computer Interaction: When a user searches for a book or document, he should be able to easily navigate through the book. Better book-readers and intuitive user interfaces need to be developed, so that the user can navigate easily. Personalization of services for easier information access is presented in [13]. Besides, the display technology for Indian language content needs to be developed and standardized, without which it would be difficult to display Indian language

text. Web browsers and softwares need to be developed with inherent Indian language support. This would enable the user to search and retrieve books in the native language without having to depend on transliteration.

Non-text Media: Digitization of non-text documents is an imminent challenge ahead. Digitization of paintings and murals is of significance and requires a completely different approach for digitization as well as utilization. Better digitization techniques need to be developed. In case of large documents such as cloth paintings, we would need a mosaicing of many images of high resolution (and small size). Efficient image processing algorithms and special information retrieval mechanisms will need to be built. Digital libraries of sculptures or three dimensional objects will need high performance 3-D scanners and fast methods and software to digitize them. Finally, audio and video libraries would be very popular, but require very robust and efficient search-retrieval expertise.

6 Conclusion

We have made considerable progress with respect to the goal of digitizing a million books. We have established semi-automatic digitization model that works very efficiently and robustly. In this process we have realized further challenges that need to be solved to realize the full potential of the activity. We have discussed in this paper the process workflow which evolved during the execution of the Million Book Project, and the performance statistics of this pipeline. We presented the challenges that were addressed from the effort towards MBP, so far. In light of the recent surge in digital library projects globally and large scale intensification of digitization efforts, we could expect almost all of man's knowledge available in digital form on the web, in the next decade or so. We could expect to see the entire cultural and historical records preserved for future generations and available across the world. Lectures, talks and presentations of teachers and researchers from the best schools in the world would be available to the students in every school, giving education and research activities a significant boost. Appreciation and exchange of art, science, music, movies, culture etc. beyond boundaries will be the real accomplishment of the "Global Village".

Acknowledgments

We would like to acknowledge Prof. Raj Reddy, CMU for his valuable guidance of this project and also for his suggestions towards this paper. We thank Prof. N. Balakrishnan of IISC-Bangalore and Prof. Rajeev Sangal of IIIT Hyderabad for their enormous support and guidance. We acknowledge the financial support received for the project from the Ministry of Communication and Information Technology, Govt. of India and National Science Foundation, USA. Technical and managerial contributions from V. Kiran Kumar and the staff at RMSC, IIIT-Hyderabad is also acknowledged.

References

1. Digital Library of India. (at: `http://dli.iiit.ac.in`)
2. Universal Library. (at: `http://www.ulib.org`)
3. Reddy, R.: The universal library: Intelligent agents and information on demand. In: ADL. (1995) 27–34
4. Million Book Project. at: `http://www.archive.org/details/millionbooks` (2001)
5. Lesk, M.E.: Understanding Digital Libraries, 2nd ed. Morgan Kaufmann, San Francisco, CA (2004)
6. Baird, H.S., Govindaraju, V., eds.: 1st International Workshop on Document Image Analysis for Libraries (DIAL 2004), 23-24 January 2004, Palo Alto, CA, USA. In Baird, H.S., Govindaraju, V., eds.: DIAL, IEEE Computer Society (2004)
7. Baird, H.S., Govindaraju, V., Lopresti, D.P.: Document analysis systems architectures for digital libraries. In: IAPR Document Analysis Systems Workshop (DAS04), Florence, Italy (2004)
8. Cole, T.W.: Creating a framework of guidance for building good digital collections. First Monday **7** (2002)
9. Baird, H.S.: Digital libraries and document image analysis. In: IAPR 7th Int'l Conf. on Document Analysis and Recognition, Edinburgh, Scotland (2003)
10. Workshop Proceedings, Tools and Resources for Digital Library, IIIT-Hyderabad, 2005.
11. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (2001)
12. Vamshi Ambati, K. Pramod Sankar, Lakshmi Pratha, C. V. Jawahar: Quality management in digital libraries. In: International Conference on Universal Digital Library, Hangzhou,P.R.China, Zhejiang University Press (2005)
13. Frommholz, I., Knezevic, P., Mehta, B., Niedere, C., Risse, T., Thiel, U.: Supporting information access in next generation digital library architectures. In Agosti, M., Schek, H.J., Trker, C., eds.: Digital Library Architectures: Peer-to-Peer, Grid, and Service-Orientation. Proceedings of the Sixth Thematic Workshop of the EU Network of Excellence DELOS, Cagliari, Italy (2004) 49–60
14. Kompalli, S., Nayak, S., Setlur, S., Govindaraju, V.: Cahllenges in OCR of devanagari documents. In: IAPR 8th Int'l Conf. on Document Analysis and Recognition, Seoul, Korea (2005)
15. Baird, H.S., Lopresti, D.P., Davidson, B., Pottenger, W.: Robust document image understanding techniques. In: 1st ACM Hardcopy Document Processing Workshop (HDP 2004), Washington, DC (2004) 9–14