The Digital Library of India Project: Process, Policies and Architecture

Vamshi Ambati[1], N.Balakrishnan[2], Raj Reddy[1,]
Lakshmi Pratha[3], and C.V. Jawahar[3]

[1] Carnegie Mellon University,
PA, USA
vamshi@cmu.edu, rr@cmu.edu
[2] Indian Institute of Science,
Bangalore, India
balki@serc.iisc.ernet.in
[3] International Institute of Information Technology,
Hyderabad, India
lakshmipratha@iiit.net, jawahar@iiit.net

Abstract

In this paper we share the experience gained from establishing a process and a supporting architecture for the Digital Library of India (DLI) project. The DLI project was started with a vision of digitizing books and making them available online, in a searchable and browseable form. The digitization of the books takes place at geographically distributed locations. This raises many issues related to policy and collaboration. We discuss these problems in detail and present the process and workflow that is established to solve them. We also share the architecture of the project that supports the smooth implementation of the process. The architecture of the DLI project has been arrived at after considering factors like high performance, scalability, availability and economy.

Keywords: digital library, digital library architecture, Digital library project of India, Universal digital library, DLI process

1. Introduction

Digital Libraries have received wide attention in the recent years allowing access to digital information from anywhere across the world. They have become widely accepted and even preferred information sources in areas of education, science and others (Bainbridge, Thompson, & H.Witten 2003). The rapid growth of Internet and the increasing interest in development of digital library related technologies and collections (McCray & Gallagher 2001) (Marchionini & Maurer 1995) helped accelerate the digitization of printed documents in the past few years.

With a vision of digitizing a million books by 2008, the Digital Library of India (DLI) project aims to digitally preserve 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, research and also for appreciation by our future generations. Ever since its inception in November, 2002 operating at three centers, the project has been successfully digitizing books, which are a dominant store of knowledge and culture. We now host close to one tenth of a million books online with about 33 million pages scanned at almost 30 centers across the country. The scanning centers include academic institutions of high repute, religious and government institutions.

In such a highly distributed environment establishing a notion of collaborative effort and distributing discrete chunks of work while maintaining uniform standards becomes a high priority task. Traditionally, digital libraries work in a closed environment and contain the process information and the content in a local repository. Although doing so increases the ease of server management and administration along with simpler resolution of process oriented issues, such an isolated set up does not scale up easily or promote collaboration across geographically distributed points of operation. This adds unacceptable delays in the implementation of the project. In projects like the DLI with such ambitious missions, a distributed environment becomes a requisite. There are discrete phases and chunks of work, which need not be collocated for operation. For example the process of scanning books can take place at one place while the image processing and web-enabling of the same could occur at a different place. Also with some planning both these tasks could go on concurrently at different places on different consignments of books in turn yielding a higher throughput. In the DLI project, we have established a flexible, yet cohesive, process that automates the entire workflow in a distributed environment. In doing so we have confronted a few problems and issues (Sankar *et al.* 2006) starting from the selection of books for digitizing, operating and establishing a protocol for being free from effort duplication, producing digital output of good quality and preservation of the digitized book objects for access in a user friendly, reliable and highly available manner.

In this paper we describe the experience gained from establishing a process involving an efficient workflow and effective policies and deploying a scalable, distributed architecture for the Digital Library of India project. We suppose that the same can be successfully applied to other similar digital libraries and digitization systems. We also attempt at throwing light on some unforeseen problems in the digitization process and issues of collaboration in a distributed environment.

The rest of the paper is organized as follows. In section 2, we give an overview of the project and its organization. In section 3, we discuss the problems and challenges that we experienced in the course of the digitization and web-enablement of books. In section 4, we discuss the process established in the project that has helped us address a few of the aforementioned problems. Section 5, discusses the architecture that addresses the issue of reliable web access of digitized content and supports the DLI process and we conclude in section 6.

2. Overview of the Project

The Digital Library of India project was initiated in the year 2002, with motivations from the Universal Digital Library project[1]. The project currently digitizes and preserves books ,though one of the future avenues is to preserve existing digital media of different formats like video, audio etc. The scanning operations and preservation of digital data takes place at different centers across India, Regional Mega Scanning Center (RMSC). The RMSCs themselves function as individual organizations with scanning units established at several locations in the region. Responsibilities of a RMSC include regulating the processes of procuring or collecting the books, distributing across scanning locations maintained by it, gathering back the digitized content from the contractors operating at those locations and hosting the same. Hence the DLI project is a congregation of RMSCs, operating parallely and independently at distributed regions across India.

The major responsibilities of the management at the DLI are to monitor the progress of the RMSCs and supply the resources necessary for its operation. We also have a contractor team which complements the setup at an RMSC. The contractor team comprises of a set of contractors who operate in the scanning locations maintained by the RMSC and have a trained personnel to execute the scanning and image processing operations. In such a highly decentralized and distributed environment, the DLI project has evolved and has been successfully producing high quality digitization of books. We now host close to one tenth of a million books containing about 33 million pages. The books come from about 15 different languages and belonged to 40 varied subjects.

3. Problems and Challenges

[1] http://www.ulib.org

In this section we present a few important operational and policy related problems and challenges that we experienced in the project. There are several appreciated research challenges in the area of digital libraries like information retrieval, multi lingual support etc, but these are not the scope of this section.

3.1. Procurement of Books

The motivation in the DLI project has so far been to preserve the rich and affluent culture and heritage of India, that is only captured in the paper and book media. As we are stepping a few milestones we realize that user profiling and usage statistics will turn out to be a promising factor in the success of any digital library. Within the number of books that have so far been enabled for online usage, we discovered from the usage logs that about 80% of the books are not accessed most of the time. Hence, considerable attention needs to be paid towards identifying and procuring books that are useful to broader communities of people. Yet another problem is that many people do not agree to the fact that digitization is the only way of preserving the books. Some are apprehensive about the possible rough use and the resultant damage to the books by the scanning centre staff particularly for the old books and palm leave manuscripts. Convincing them to loan such rare pieces of information for scanning is important, if the DLI has to contain useful collection of books.

3.2. Incomplete and Incorrect Metadata

Most of the books scanned in the DLI project are procured from sources like libraries and government archives and hence contain metadata entered by knowledgeable personnel which can be relied upon, but is still debatable due to individual biases. However a major portion of the sources of books in the project have metadata only in non-digital formats and so these have to be fed manually. This process though inevitable is understood to be prone to errors. Due to this varied sources of book flow in the DLI project in multiple languages and due to the lack of standard formats, metadata is missing, incorrect or incomplete or sometimes difficult to interpret. Inaccurate metadata hinders fruitful search and retrieval of books, categorization and at the same time brings in scope for duplicate entries of the same book.

3.3. Duplication

As mentioned earlier, due to the varied sources of books, like libraries, government organizations, institutions and personal collections that are distributed across various parts of the country, duplicates could arise between scanning locations maintained by an RMSC and also across different RMSCs. Effort put into scanning a book, processing the images and quality assurance can not be afforded to be spent on duplicates. Communicating metadata across centers and within scanning locations is important. The Duplication of the books can be identified only using metadata of a book like the title, author, publishing year, edition, etc. However, if the metadata is incorrect, missing or incomplete as discussed in the previous section, it makes the duplicate detection all the more difficult.

3.4. Data Management

Assembling the data and making it available for easy access is one of the most important phases of any digitization project (Ingo Fromholz 2004). Each Mega scanning centre is responsible for gathering the metadata and the scanned content from the contractors operating at its scanning locations. This data is to be enabled on the web and also preserved for future. Enabling many tera bytes of data for access to everyone in a higly reliable manner is needed for the success of the efforts put into the digitization process. Also data synchronization and management across centers needs to be done to reduce duplication and ensure reliable high availability and immediate recovery in the event of storage media failures and server failures. Finally, digital preservation of the collections for a long future still remains a very significant problem faced by any digital library (Barroso, Dean, & Urs Holzle 2003).

4. Process and Workflow

Most of the problems mentioned in the above section have been addressed by adhering to a rigid process and by establishing a scalable and interoperable software architecture. The process and the workflow includes several committees with individual responsibilities. In this section we first talk about the categorization and storage schema of metadata that is used at DLI, then discuss the workflow, the committees and the policies that have assured a high quality output from the DLI.

4.1. Metadata

The process at DLI is metadata centric. Every book that is scanned and stored is associated with metadata for identification and search and retrieval. Identifying the metadata that should be preserved along with the digital objects is a debatable topic. At DLI project we had several discussions as to what is the right schema for metadata of a book and finally narrowed down on the following three sub-categorizations

1. Regular Metadata:
   Regular metadata contains information about the book like title, author, date of publication, publisher, ISBN, keywords, subject, language etc. We follow the widely understood and accepted Dublin core[2] format and extended with a few extra fields like edition information of the book etc. This metadata primarily helps us to identify, categorize and retrieve the book.
2. Administrative Metadata:
   Administrative details of the book, like the location of scanning of a book, the source of the book, details of scanning of book etc may not be of interest to the book readers but are useful to the operational organization. It can be used to trace the progress of the project, generate reports and identify bottlenecks in scanning process etc. For example, we could trace the scanner producing low quality scans etc.
3. Structural Metadata:
   We have adapted the structural metadata concept proposed in (Dushay 2002) for a book object in our digital library. This metadata contains information pertaining to each page like the size of each page, whether the page is blank, or has an important context attached to it like the beginning of chapter, end of chapter, index, preface, table of contents etc. Such information enables us to improve the navigation of the end user through the book and also improve search and retrieval systems.

Each digital book object contains all three forms of metadata. Regular metadata is entered by source librarians or the contractor hired librarians before digitization begins. Admin metadata is gathered at the contractor end during the process of digitization. Structural metadata is manually entered by the contractor and also automatically detected by learning techniques, although the heterogeneity of the structure of pages in a book prevent complete automated detection.

4.2. Workflow

Procurement team identifies the books to be digitized. The books are then directly shipped to various scanning locations operated under an RMSC. The digitization of a book starts with an expert librarian entering the regular metadata for the books that need to be scanned. The metadata is first uploaded onto the DLI portal hosted at the RMSC for checking of possible duplicates from else where at other scanning locations. However due to a continuous flow of books from libraries all over, a significant overlap is expected not only between scanning locations but also across RMSCs. Hence the uploaded metadata has to be synchronized with the other RMSC databases and then duplicates are detected in the

[2] http://dublincore.org/

uploaded metadata. This ensures prevention of duplicates in the system,

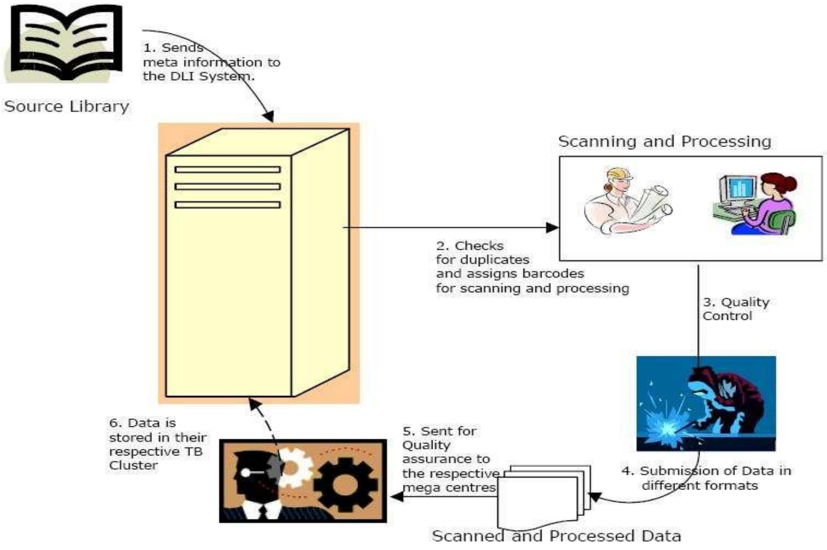
Source Library
1. Sends meta information to the DLI System.
Scanning and Processing
2. Checks for duplicates and assigns barcodes for scanning and processing
3. Quality Control
4. Submission of Data in different formats
Scanned and Processed Data
5. Sent for Quality assurance to the respective mega centres
6. Data is stored in their respective TB Cluster

Figure.1. Process and Work flow in the DLI Project

assuming the metadata is legitimate. Books are then digitized by the contractor and given back to the RMSC under which he operates. The product is tested for quality standards and approved for upload onto the servers.

4.3. Team Organization

The following three teams play a major role in the workflow of DLI.

1. Procurement Team

Identifying and procuring rare monumental works is the responsibility of the procurement committee. It approves books that are useful for a wider range of communities. Usage statistics of the already online books are collected from the system log files and given for the inspection of the procurement committee. The procurement committee uses them to decide upon the books to be procured and scanned. It is also responsible for handling issues related to copyright and intellectual property of the books.

2. Metadata team:

This team consisting of librarians and technicians 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 by remotely distributed librarians who can log in to the system and make necessary corrections over the web. Automated identification and correction of metadata using statistical techniques and human feedback has also been yielding successful results.

3. Quality Assurance team:

This team verifies the digitized content returned from the contractors and approves for uploading and hosting on the web (Ambati et al 2005). They perform the check for duplicates, improper scans, damaged pages, missing pages, file formats and a few other parameters to ensure that the quality standards defined in the DLI are met. Administrative issues regarding the decision making of the undefined errors found in the digitized books and content is also made by this team. The team also ensures the process is carried out in the defined manner and performs process audits for applying the improvement strategies.

4.4. Policies

The following are a few policies followed in the DLI project to improve the process and workflow in turn improving the quality of the product:

1. Peer reviews:
   Peer reviews are quite useful in research and educational purposes and is also adopted as a good practice by the source software engineering community. Peer reviews are conducted amongst the contractors during which the contractors and the management sit together to discuss and critique constructively on their work, responsibilities and the output produced.
2. Expertise exchange:
   To be up to date with the results of research and improving technology, the tools and resources that are developed at various centers and academic institutions are shared periodically. Workshops and training sessions are conducted discuss the tools and to train the personnel concerned.
3. Process Reviews:
   Project personnel involved in the decision making process, meet to conduct process reviews and debug the process based on feedback from the various active teams and committes.

5. Architecture of DLI Project

In this section we describe our architecture that supports the process and the workflow discussed in the earlier sections. The architecture of the DLI project is similar motivated by factors like scalability, ease of maintenance, dependability and economy. All the tools and technologies used in the DLI are free software. Many issues like interoperability, collaboration arise due to the multitude of books and languages that are scanned at the various scanning locations and the differences in the infrastructures used top reserve these digital objects. We solve this by deploying a distributed decentralized architecture for the DLI project and by modularizing the tasks, using technologies like XML, Databases, Web services etc. First we talk about the architecture of the DLI portal hosted at each Mega center (DLI-RMSC) and then propose an architecture for organizing these individual portals in a decentralized and service oriented manner to ensure a highly available and dependable DLI system.

5.1. Architecture of DLI at a Regional Mega Scanning Centre

Each Mega centre hosts the books that are scanned in the locations maintained by it. Currently there are three operational mega centers. The architecture adapted by each RMSC is similar to the one shown in Figure 2. 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 Linux and enhanced by LTSP[3],an add on package for Linux that supports 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 replaced at a nodes in the cluster instantaneously without the need for operating system installations and configurations. We have customized the kernel in LTSP to support hard disk recognition and usb hot plug, and to run a light weight Apache web server.

As shown in the Figure 2 the 'Linux Loader' machine runs a copy of this distribution of the Linux with LTSP. Each data server in the data cluster downloads the kernel over the private intranet and boots from it. The servers implement a hardware based 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

[3] http://ltsp.org/

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, modularizes 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. Along with the metadata of the book, the database also contains pointers 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 requests, from the particular server in the cluster and is broadcast to the user. The ’proxy servers layer’ between the Data cluster and the DLI-RMSC portal also has a caching mechanism enabled that handles repeated requests to the book pages and ensures quick response times. Books are also preserved in text format which is searchable. The search is now only limited to books in English language, due to non availability of optical character recognizers for other languages. The search is supported by Lucene[4].

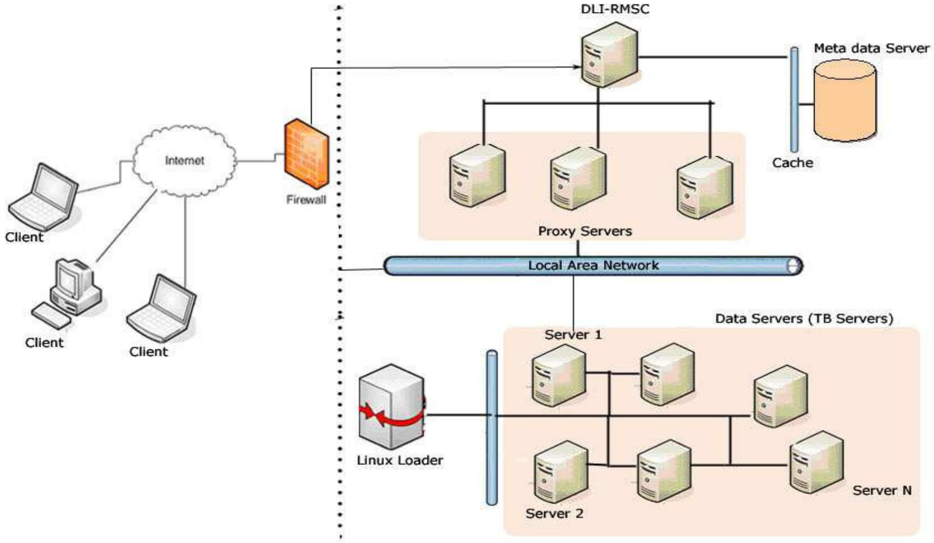

Figure.2. Architecture of the DLI hosted at a Regional Mega Scanning Centre

5.2. Distributed Architecture of the DLI Project

The scanning operations of DLI take place at different locations in which the RMSC operates and the digital data from its region is accumulated to be hosted online. The DLI as such follows a distributed and decentralized architecture with each RMSC as an independently operating node. Decentralized architectures by definition avoid having central points, as they are candidate single points of failure and a performance bottleneck. However, since the digitization process is a cumbersome process, the data that is the end product of the process is very sacred. Hence redundancy is always advised in a data centric project like DLI and therefore every RMSC which is a node in the decentralized architecture of DLI hosts the complete data from the other nodes. Synchronization of complete book content between nodes is currently by physical transfer of Terabytes of information between the RMSCs.

[4] http://lucene.apache.org

We propose a Service Orientated Architecture (SOA) (see Figure 3), for smooth interaction between the nodes in the decentralized architecture. The core architecture is motivated from that of Google (Rothenberg 1999), and we adapt and extend the same to suit the requirements of DLI. The following are a few advantages of the decentralized and SOA architecture of the DLI:

- Web services address issues of interoperability that arise due to varying media, databases, languages, middleware and operating systems across RMSCs.
- Metadata of books is synchronized between RMSCs via web services, on a periodic basis. This also helps in duplication verification across RMSCs.
- Other specific features like copyright information verification, statistical reports etc can also be exposed by the RMSCs and can be utilized across the DLI project via web services.

A user can login to the central site and request to read books online at which point he is redirected to one of the closest RMSCs and be served from there. This ensures quick response times for the user and also reduces to some extent load on any one set of servers.

6. Conclusions

In this paper, we have discussed a few issues and problems faced in running a large scale digitization project like the DLI and shared our experiences in handling them. Also we have given an overview of our process, policies and workflow and described the architecture that ensures a smooth deployment of the process.

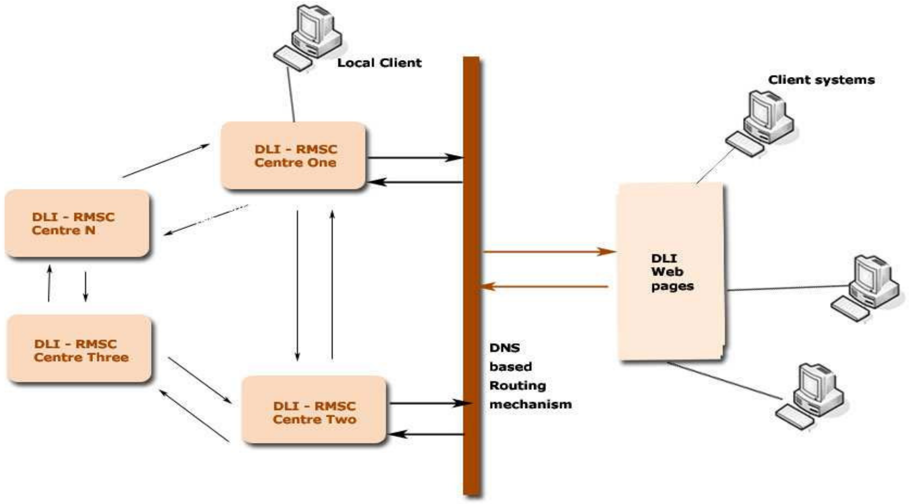

Figure.3. Decentralized SOA based Architecture of DLI

References

Ambati, V.; Sankar, P.; Pratha, L.; and Jawahar, C. 2005. **Quality management in digital libraries**. In *In Proceedings of 1st ICUDL*.

Bainbridge, D.; Thompson, J.; and H.Witten, I. 2003. **Assembling and enriching digital library collections**. In *Proceedingsof the 3rd ACM/IEEE-CS joint conference on Digital libraries*, 323–334.

Barroso, L. A.; Dean, J.; and Urs Holzle. 2003. **Web search for a planet: The google cluster architecture**. *IEEE Micro,* Volume 23(2) :22–28.

Dushay, N. 2002. **Localizing experience of digital content via structural metadata**. In *JCDL '02: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, 244–252. New York, NY, USA: ACM Press.

Ingo Fromholz, Predrag Knezevic, et al . 2004. **Supporting information access in next generation digital library architectures**. In *DELOS Workshop: Digital Library Architectures*, 49–60.

Marchionini, G., and Maurer, H. 1995. **The roles of digital libraries in teaching and learning**. *Communications of the ACM,* Volume 38(4):67–75.

McCray, A. T., and Gallagher, M. E. 2001. **Principles for digital library development**. *Communications of the ACM* , Volume 44(5):48–54.

Rothenberg, J. 1999. **Avoiding technological quicks and finding a viable technical foundation for digital preservation**. *Rep. to Council on Library and Information Resources.*

Sankar, K. P.; Ambati, V.; Pratha, L.; and Jawahar, C. V. 2006. **Digitizing a million books: Challenges for document analysis**. In *Document Analysis Systems*, 425–436.