TCDL Bulletin
Current 2006
Volume 3 Issue 1

Digital Library of India

A Testbed for Indian Language Research

N. Balakrishnan
Supercomputer Education and Research Center
Indian Institute of Science
Bangalore, 560012 India
<balki@serc.iisc.ernet.in>
<http://swati.dli.ernet.in/balki/>

Raj Reddy, Madhavi Ganapathiraju and Vamshi Ambati
School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15213 USA
{rr, madhavi@cs.cmu.edu}, <vamshi.ambati@gmail.com>

Abstract

This paper describes the goal of the Universal Digital Library Project (UDL) and presents the approach taken by – and the technological challenges associated with – the Million Books to the Web Project (MBP). The Digital Library of India (DLI) initiative, which is the Indian part of the UDL and MBP, is discussed. DLI fosters a large number of research activities in areas such as text summarization, information retrieval, machine translation and transliteration, optical character recognition, handwriting recognition, and natural language parsing and morphological analyses. This paper provides an overview of the activities of DLI in these areas and shows how DLI serves as a multilingual resource.

Introduction

The recent advances in computer, storage and communication technologies are so stunning that it

prompted one of the authors of this paper (Raj Reddy) to envision and explore the possibility of storing in digital form all the knowledge ever produced by the human race and making this content available free of charge to be browsed and searched by anyone, anywhere and at anytime. This vision is the goal of the Universal Digital Library Project (UDL). In addition, Reddy envisions a future where almost all information will be readable by humans as well as machines. The trend would be such that any information that is not on-line and accessible to search engines may become unusable. In a thousand years, only a few of the paper documents we have today will survive the ravages of deterioration, loss, and outright destruction. Hence there is an urgent need to preserve our knowledge and heritage in digital form.

As a part of Raj Reddy's grand vision, a mission to digitize one million books was embarked upon as a collaborative project involving many countries, especially India, the United States and China. In this paper, we present the approach taken and the technological challenges associated with the Million Books to the Web Project (MBP).

An interesting offshoot of our efforts to make knowledge available free of charge to everyone is the opening up of research opportunities in language technologies – particularly for Indian languages – in order to make sure the language in which that knowledge exists does not become a barrier to information access. Language technologies research in Indian languages has so far been impaired by the lack of resources pertaining to Indian languages in the form of text or speech corpora. Compared to the corpora in English and other European languages, or in Chinese, the resources available for Indian languages have been very limited. This situation is being changed by the Digital Library of India (DLI) initiative, which is the Indian part of the UDL and MBP.

DLI today has scanned over 289,000 books composed of approximately 105 million pages in the Indian and English languages. Digital representation and storage mechanisms have been developed for Indian languages, and a large number of applications are being built to store, process, retrieve and present the Indian language content. The Digital Library of India fosters a large number of research activities pertaining to language technologies for Indian languages, and acts as a testbed for developments made in areas such as text summarization, information retrieval, machine translation and transliteration, optical character recognition, handwriting recognition, and natural language parsing and morphological analyses. We present here an overview of the activities of DLI in these areas and show how DLI is acting as a multilingual resource, even without the availability of manually curated data.

Vision: Digitization of all human knowledge

For the first time in history, technology seems to favour the possibility of digital preservation of all the significant literary, artistic, and scientific works of mankind, as well as the potential of free access to them from every corner of the world. A Universal Digital Library (UDL) has the potential of improving the global society in ways beyond measurement. The Internet can house a Universal Library that is freely accessible to everyone. This would revolutionize education for all our future generations. There were about 10 million unique book and document editions before the year 1900, and about 100 million since the beginning of recorded history. An average-sized book is around 250 pages and would require about 50 MB of disk storage if the book were stored as compressed images. Thus, all the books and documents ever produced by the human race would require 5 peta byte of storage. Even if we multiply this by a factor of 200 for all other forms of knowledge, such as music, images, audio and video, the

total of that information could be stored in a zeta byte server. With the storage capacity of digital disks increasing by a factor of 1,000 in ten years, it looks technically feasible and financially affordable to articulate the vision to store on the computer all forms of knowledge ever produced by the human race. With new digital technology, though, this task is within the reach of a single concerted effort for the public good, and this effort can be distributed to libraries, museums, and other groups in every country. This formed the motivation for the grand vision of the Universal Digital Library Project.

Mission: A million books on the web by 2008

It is believed by some that the goal of creating a Universal Digital Library is impossible and that attempting to create it could take hundreds of years and still may never be completed. Nevertheless, as a first step toward realizing this grand vision, a project was proposed that would create a Universal Library starting with a free-to-read, searchable collection of one million books available to everyone over the Internet by the year 2008. This first major project toward building a Universal Library is named the Million Book Digital Library Project (MBP). Within 10 years, it is expected that the collection will grow to 10 Million books. The result will be a unique resource accessible to anyone in the world, 24x7, without regard to nationality or socioeconomic background. Typical large high school libraries house fewer than 30,000 volumes. Most libraries in the world have fewer than a million volumes. The total number of different titles indexed in OCLC's WorldCat is about 55 million. A library of one million books, therefore, contains more than the holdings of most high schools, and is equivalent in number of volumes to the number in the libraries at many universities, representing a useful fraction of all available books.

The MBP is a multi-country project, and that part of the project being performed in India is called the Digital Library of India (DLI). A secondary objective of the MBP project, and hence the country-specific Digital Library exercises such as DLI, will be to provide a testbed that will support other researchers who are working on improving the technology required for scanning and indexing. The corpus the MBP project creates will be one to three orders of magnitude larger than any existing free resource. It is expected that when DLI matures, more than 10,000 books will be available in each of the major Indian Languages. Such an immense and unequaled volume of data would also provide an excellent resource for language processing research in areas such as optical character recognition (OCR), machine translation, summarization, intelligent indexing, and information retrieval.

Collaboration

Accessing, scanning and web hosting a million books is a major logistical challenge. A million books can be expected to comprise approximately 250 million pages. Throughput of a scanner used in the project (Minolta PS7000) is about 6,000 to 10,000 pages per day, based on three-shift operation. Using 100 scanners, and a 1,000-person workforce, the task of scanning a million books could be completed in about 420 working days, or roughly two years of three-shift operation. However, given the uncertainties of having skilled workers available at all times, and experiencing equipment failures and power outages, it is more reasonable to expect that the scanning would be completed in around four years.

The United States of America, India and China are currently the major collaborators in the MBP project with regard to technology development, operations and management. Collaboration among these countries has been a factor in the project's success so far, because these three countries provide access to

a number of large academic libraries, of which over 700 are OCLC members from India and China. Also, since the scanning centers can be established anywhere, UDL has chosen to establish them close to these academic libraries so that the books can be transported with minimal difficulty.

The collaboration so far has been quite fruitful, with more than a few thousand books digitized every month from all the various scanning points established in India and the US. Currently, we have more than 200,000 books scanned and ready for use.

The Approach: Digital Library of India

The Indian Institute of Science (IISc), Carnegie Mellon University (CMU), the International Institute of Information Technology, Hyderabad (IIITH) and many other academic, religious and government organizations, totaling about 21 "Content Creation Centres", have become partners in the Digital Library of India (DLI) initiative for the digitization and preservation of Indian heritage present in the form of books, manuscripts, art and music. Each centre brings its own unique collection of literature into the digital library. Many authors have cooperated by contributing their books to the digital library and making them available free of charge to anyone. This digital library is also intended to be a testbed for Indian language Research. DLI is intended to be a leading and contributing partner in worldwide efforts toward making knowledge free. A pilot project to scan around 10,000 books was initiated at CMU and then followed up at IISc, IIITH and other organizations; all the processes involved have been perfected. The vision is to use the disruptive technologies like the ICT to preserve all the knowledge of the human race in digital form and make that content searchable, independent of language and location, and to ensure that the rich cultural heritage of countries like India is not lost during the transition from paper to bits and bytes, as they were lost during a former transition of cultural content from palm leaves to paper.

Today, most of the works created by humans – be they in the form of books or music or movies – are "born digital". Hence if we build the digital library with a proper framework and architecture, it should become possible to bring the library up to date more easily than it was to digitize physical formats. But for the time being, we are concentrating on digitizing these physical formats. For the digitization of old books, we have developed a complete process comprised of scanning with planetary scanners, cropping, cleaning up images, and using software for the OCR conversion of English documents, format conversions and search engines. With available technologies, a 500-page paper book can be digitized and made available on the web in about two hours without having to unbind the book. So far, more than 289,000 books have been scanned, of which nearly 170,000 are in Indian languages. More than 84,000 books (25 million pages) are available on the DLI web site hosted by the Indian Institute of Science (http://www.new.dli.ernet.in), and more than 149,000 books (43 million pages) are available on the DLI web site, which is hosted by the International Institute of Information Technology (http://dli.iiit.ac.in). The link to other partner sites are also provided through (http://www.new.dli.ernet.in). Contents between the two sites overlap in order to ensure fail safe availability. The books can be accessed from either of these web sites. With the success of the joint efforts of the IISc-CMU collaboration, many other nations including China and Egypt have shown interest in participating in this effort, making it truly a global effort in knowledge sharing. China has made significant progress and has taken the Million Books to the Web as a National Initiative.

Technological Challenges

Quality Management

In large digitization projects, with collaborative work and distributed efforts of various parties involved in the process, we often find a compromise of quality, which enables errors to creep in. As is the case in any digitization project, DLI work is performed by humans and machines. Hence, occasional errors are possible, and these can be broadly classified into two categories: human errors and machine errors, both of which we discuss below.

Human Errors

Human errors in the digitization process are perhaps the costliest of the errors; they arise most often due to miscommunication between project staff or to staff incompetence or non-adherence to process and standards. Most of the books scanned in the DLI project are procured from sources like libraries and government archives, and the records for these books contain metadata entered by knowledgeable personnel. Although, in general, the metadata can be relied upon, the quality of that metadata are nevertheless subject to individual biases. In addition, for a major portion of books scanned in the project, accompanying metadata exists only in non-digital formats, and these have to be entered manually. For the data entry of metadata of a book, we largely rely on librarians to assess the accuracy and credibility of that metadata. But this situation is not foolproof. For example, a particular librarian might not be well advised about the hierarchy and ontology of book classification and so might classify a book as belonging to the category of "Art" when it really belongs in the category of "Music". Consequently, the misclassified book would not show up in search results for the end-user who is interested in and searches for "Art", and instead would show up as a non-relevant search result to a person searching for "Music". An even worse situation would arise if the book were erroneously assigned to a completely irrelevant category.

Similarly, the process of scanning and producing digital content needs to be done with proficiency and care. An improper scanning operation made without following the standards set, may result in a digital collection that is not useful or suitable for an end-user. These problems manifest in various forms like the page of a book slipping while being scanning or an incomplete scan of a page. Manual errors can incur significant costs to the project, as firstly there is no satisfactory way to identify such errors when data is generated on such a massive scale, and secondly the erroneous data generated will not useful to end-users, thus undermining the purpose of the project.

Machine Errors

Machine errors are those that creep in due to the understandable limitations of the software being used or from improper configurations of the machines and software. Data generation, which is done during the scanning, is the key phase of the digitization process as we obtain digital images from the books. Once these images are obtained, the book is sent back to the library from which it came, and hence any problems that require that the book be re-obtained and re-scanned are costly errors. For example, an image processing algorithm that checks for the 'skew' aspect in images may have certain limitations for detecting it and could classify a non-skew image as containing skew. Such a problem could be rectified with better versions of the software whenever available. However errors due to improper configuration

of scanners during the scanning phase are more serious, as the data generated becomes less useful due to its low quality.

Optical Character Recognition (OCR) is the software module that reads an image, understands the textual content in the image and outputs the text. Text is the appropriate means for storing data, occupies less storage than images do, is easily editable, and helps in the indexing and searching of the documents. OCR is thus a very crucial aspect of the process of creating the digital library. However, OCR is not 100% accurate, due to the various limitations imposed by the underlying recognizer. Let us assume that a page consists of 30 lines, each containing 30 words, and that a word, on average, consists of 4 characters. In that case, a typical page would contain 3600 characters. If the OCR is 99% accurate, it will result in 36 erroneously converted characters per page, and therefore for a book of 500 pages there would be about 18,000 errors in the OCR-generated text. Such a situation would be unacceptable to the end-user, although it does not significantly affect the efficiency of indexing and searching.

Data management

Scalable and sustainable architecture

Assembling the data and making it available for easy access is one of the most important phases of any digitization project [1]. Each Mega scanning centre is responsible for gathering the metadata and the scanned content from the contractors operating at the 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 highly 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 time into the future remains a very significant problem faced by any digital library [2].

Preservation management

The books scanned for the DLI project are for the use of everyone for the foreseeable future. Hence preserving the content of these books is important. In addition, the data being frequently commuted between the centres needs to be preserved uniquely to ensure easy workflow management. Every book that is scanned and stored is associated with a unique barcode and descriptive metadata for identification, search and retrieval.

Synchronization across different centers

Because the books to be scanned come from various sources – like libraries, government organizations, institutions and personal collections – that are distributed across the country, there could be duplicates among scanning locations maintained by a Regional Mega Scanning Centre (RMSC) and also across different RMSCs. However, the project cannot afford the extra cost of scanning these duplicate books, processing their images, and performing quality assurance on them. Thus, communicating metadata across centers and within scanning locations is important. The duplicate books can be identified only by 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.

Rich metadata

The metadata formats that are traditionally used for physical books, though comprehensive, are not sufficient for handling digital objects. Hence, we have had several discussions in the UDL and DLI projects about identifying the metadata that should be preserved along with the digital objects, and we finally narrowed down our requirements to the following three sub-categories of metadata:

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 metadata format, extended with a few fields like edition information of the book, and the use of OM transliteration for Indian language texts developed for DLI.
2. **Administrative Metadata**: Administrative details of the book, like the location where the book was scanned, the original source of the book, scanning details, etc., may not be of interest to the end-users of the book but are useful to the operational organization. These details can be used to trace the progress of the project, generate reports and identify bottlenecks in the scanning process. For example, it would enable us to trace the scanner that was producing low quality scans.
3. **Structural Metadata**: We have adapted the structural metadata concept for a book object in our digital library. This metadata contains information pertaining to each page, like the size of each page and whether that page is blank or has an important context attached to it – such as the beginning of chapter, end of chapter, index, preface, table of contents, etc. Such information enables us to improve the navigation of the book for the end-user and also improves search and retrieval systems.

Replication of storage

Preservation of digital data is known to be a hard problem to solve. We attempt to address this problem to the best possible extent, by replicating the resources. All the books scanned in the DLI project are replicated and preserved in different locations across the world. We also store the content in two different formats – DVD/CD and hard drives. We are currently planning to have 3 different locations across the world (CMU, IISc and IIITH) where all the data scanned and digitized has been preserved in the above mentioned formats.

Architecture of the ULIB Project

In this section we describe the architecture that supports the process and workflow discussed in earlier sections of this paper. The architecture of the DLI project is motivated by factors like scalability, ease of maintenance, dependability and economy. All the tools and technologies used by DLI are open source. Many issues related to interoperability and collaboration arise due to the huge number of books in different languages that are scanned at the various scanning locations, and the differences in the infrastructures used to preserve these digital objects. We solve these issues by deploying a distributed, decentralized architecture for the DLI project and by modularizing the tasks, using technologies like XML, databases, web services, etc. Below, we first describe the architecture of the DLI portal hosted at

each Mega center (DLI-RMSC), and we 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.

Architecture of Mini UDL and DLI hosted at Mega centre

Each 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 Regional Mega Scanning Centre (RMSC) is similar to the one shown in Figure 1.

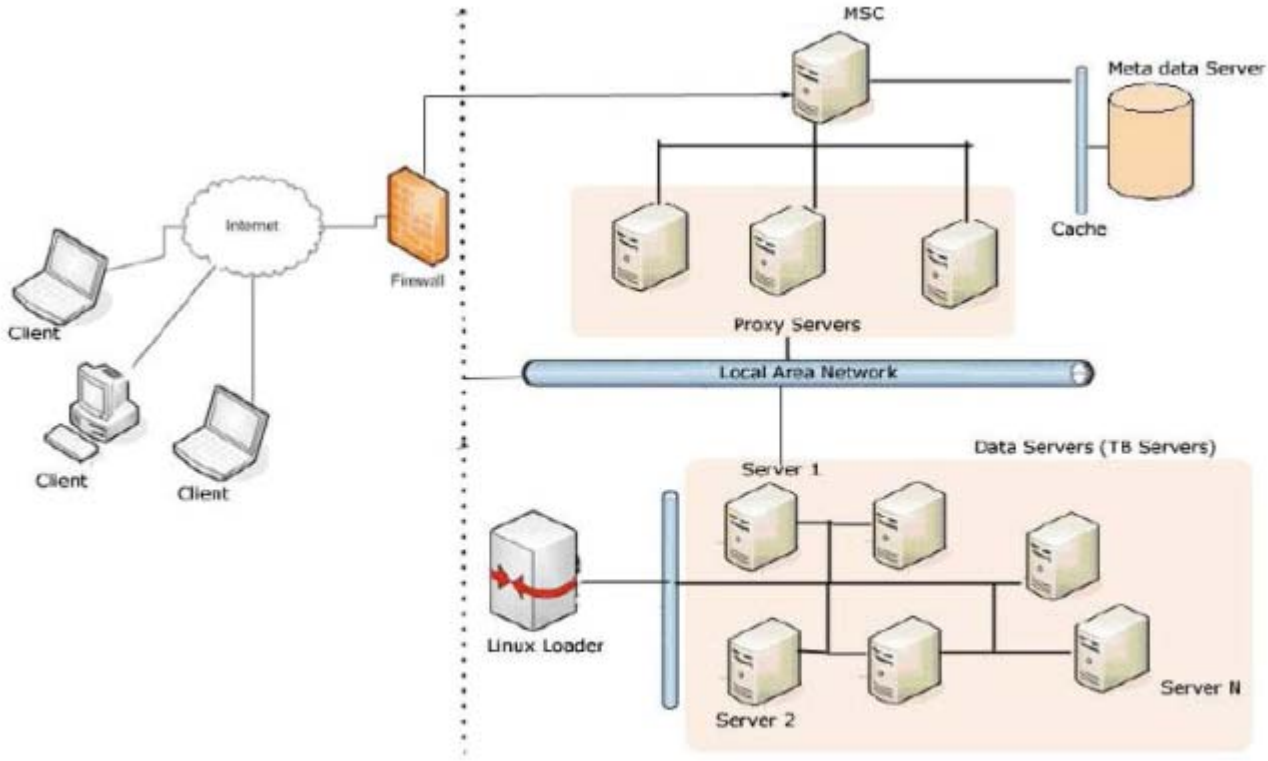

Figure 1. Architecture of the UDL and DLI hosted at a Regional Mega Scanning Centre (RMSC)

The digital objects produced by the scanning 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 LTSP3, 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 also easy to manage, such that we can add or replace data 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 hotplug, and to run a lightweight Apache web server.

As shown in Figure 1, 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. In addition, for data restoration in the event of an irrecoverable crash, a redundant copy of the complete data is present on external storage media. The 'metadata server' is a repository of the complete metadata, which is in XML. XML has been chosen for its important role in enabling 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 the RMSC and contractors, and it 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 user interface that a user can log onto, use to query the metadata, and retrieve books he or she 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 viewing the complete content of the book, the location of the book in the data cluster is gathered from the database, the content is retrieved over http requests from the particular server in the cluster and then 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 for the book pages, and this ensures quick response times. Like the metadata, books are preserved in text format, which makes them searchable. Currently, search is limited only to books in the English language, due to unavailability of optical character recognizers for other languages. The search is supported by Lucene.

Distributed Architecture of DLI Projects

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. 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 that is a node in the decentralized architecture of DLI hosts the complete data from the other nodes. Currently, synchronization of complete book content between nodes is by physical transfer of Terabytes of information between the RMSCs.

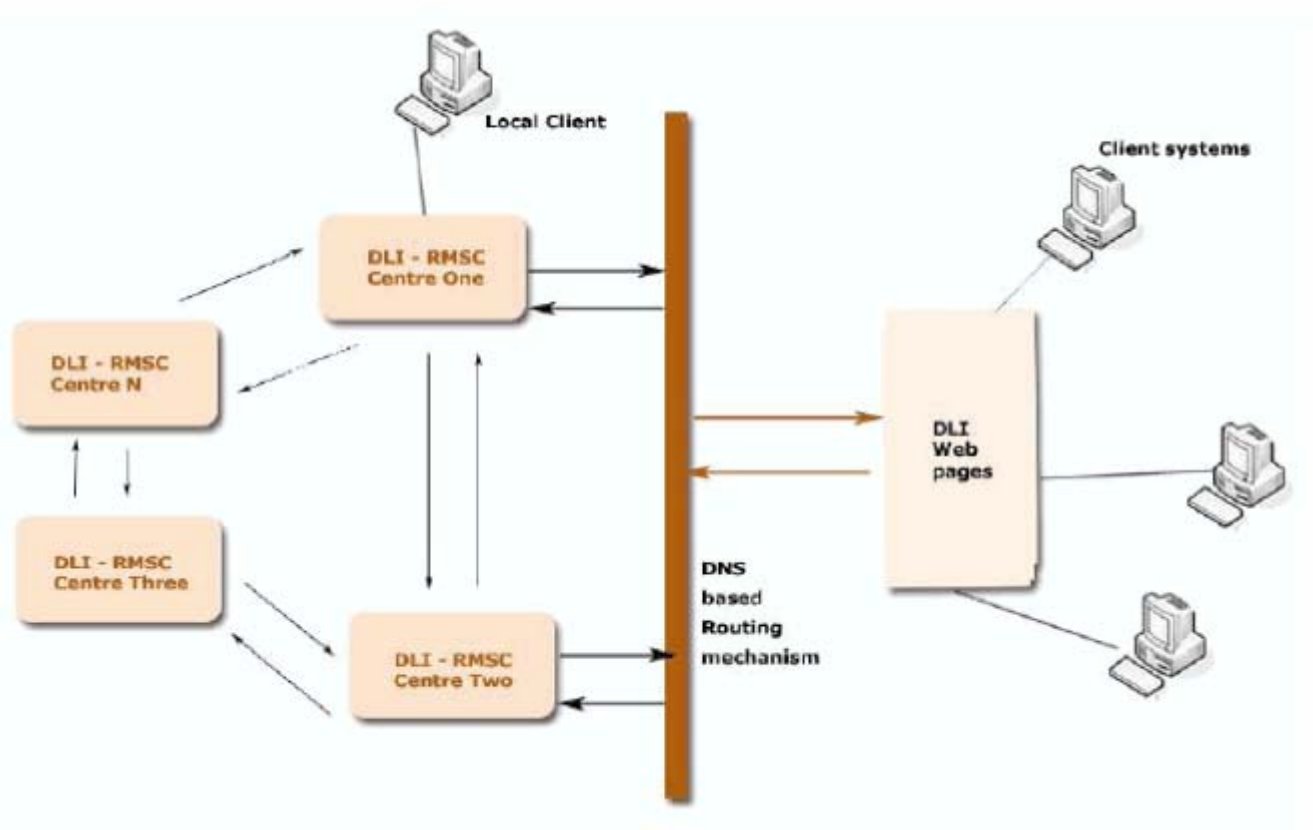

Figure 2. Decentralized SOA-based Architecture of DLI

We propose a Service-Orientated Architecture (SOA) (as shown in Figure 2) for smooth interaction between the nodes in the DLA decentralized architecture. The following are a few advantages of the decentralized SOA 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 log into the central site and request reading books online, at which point he is redirected to one of the closest RMSCs and will be served from that RMSC. This ensures quick response times for the user and also reduces, to some extent, load on any one set of servers.

Indian Language Technology Research under the DLI Project

The Million Books to the Web Project (MBP) acts today as a catalyst for research in Indian Languages. The primary goal of the DLI project, apart from making the books available online, is to make them available in fully functional form. The DLI is not simply a static repository of books – it has made possible bringing home the language and information processing technologies for Indian languages. In

the rest of the paper we describe some of the research activities progressing under the umbrella of DLI.

Major impact contributions of the Digital Library of India to the Indian language information technologies are:

1. OCR for Indian languages
2. OM transliteration, which is an integral component of all the other systems
3. a text editor for Indian languages that is available for everyone
4. the book-reader interface that extends the availability of the Digital Library from "*anyone, anytime, anywhere*" to "*any- language*"
5. a machine translation system that we call *Good-Enough Translation* (GET-across) and
6. a search engine for Indian language texts.

These are briefly described below.

OCR in Indian Languages - Kannada

Designing an accurate OCR in the Indian languages is one of the greatest challenges in computer science. Unlike European languages, Indian languages have more than 300 characters to distinguish, a task that is an order of magnitude greater than distinguishing 26 characters. This also means that the training set needed is significantly larger for Indian languages. It is estimated that at least a ten million-word corpus would be needed in any font to recognize Indian languages with an acceptable level of accuracy. DLI is expected to provide such a phenomenally large amount of data for training and testing of OCRs in Indian Languages. Many of the contents, besides scanned images, have been manually entered for this purpose. Using this extremely large repertoire of data, a Kannada OCR has been developed.

Optical character recognition for Kannada: The first block of the OCR is the segmentation algorithm that segments lines, characters and, within the character, a 32 X 32 block to identify the different key strokes that make up the character. These take care of the morphological dilation, base character, vowel modifiers and consonant conjuncts. Base characters are then normalized to 32 X 32, and the consonant conjuncts, as well as modifiers, are resized to a 16 X 16 matrix. Through a series of signal processing algorithms using DCT and KLT, the features are extracted. Structural features include aspect ratio, stroke at different orientations, and height of the segment, in the top zone, and the width of the character in the middle zone.

A neural network-based classifier is then used for training with the extracted feature vectors and testing. The current level of accuracy that we get is around 96-97% on clean documents scanned at 400 dots per inch. This accuracy falls to 40-50% if the image is of bad quality. Efforts are underway to further improve the accuracy of the OCR with better segmentation, faster speeds and enhanced training. This OCR is currently being extended to other Indian languages, including Tamil.

Om transliteration: Unified representation for Indian languages

India is fast becoming a software superpower; however, PC penetration is merely 1.4% compared to that of television (17%) and telephone (5%). One of the limiting factors in low PC usage is the

unavailability of the operational software in native Indian languages, and the language barriers between people. While the development of an operating system in a native language is one solution, this solution is likely to be limited to only a few languages. If the Indian language texts were instead available in parsable English-like texts, they would benefit from the advances in the language processing of other international languages. Isolated development of digital representations for the different Indian languages may further widen the language barrier in the country.

Thus there is a need for the development of a digital representation that lays a common foundation for all Indian languages. For seamless adaptation of algorithms in language technologies, this representation must also be parsable by universal language processing tools and algorithms, such as for machine translation, information retrieval, text summarization and statistical language modeling. The representation must exploit the similarity in the alphabet of the various Indian languages. A large number of Indians are bilingual; while they can freely read and write in the one Indian language that is their mother tongue, and English, they can understand many other Indian languages as well due to the similarity of the origin of words. With this backdrop, we first developed a representation scheme for Indian languages, called *OM transliteration*, which formed the basis for all other work in Indian language research at DLI.

Om uses the same representation for keyboard input and formation, and digital storage. It is similar to ITRANS in that it uses combinations of letters from the English alphabet to represent Indian language syllables. ITRANS is a representation of the Indian language alphabet in terms of ASCII. (http://www.aczoom.com/itrans/). However, the OM transliteration developed under DLI is case-independent, and avoids excessive use of non-alphabetic characters; where used, they are consistent. Furthermore, the English alphabet combinations are designed such that they are easy to remember at the time of input using a standard keyboard, and they are also natural to read like English. The case-independent representation allows the use of sentence and title case writing in a natural fashion; in addition, the texts are more highly readable than their ITRANS counterparts. It may be noted from any ITRANS text that the large mixture of capital and small letters and non-alphabetic characters leaves ITRANS text highly difficult to read.

Om's features enhance usability and readability; it has been designed on the following principles:

1. easy readability,
2. case-insensitive mapping (While preserving readability, this feature allows the use of standard natural language processing tools for parsing and information retrieval to be directly applied to the Indian language text.), and
3. phonetic mapping, as much as is possible. (This makes it easier for the user to remember the key combinations for different Indian characters.)

ASCII representation may be used simply as a means of typing the text with a standard keyboard. For transliteration to Indian languages, OM representation is mapped to the Indian language fonts for display or converted to any other format, such as Unicode, where required. When a user is not interested in installing language components, or when the user cannot read native language script, the text may be read in the English transliteration itself. Because India is a multi-lingual country with an inter-mixed population, often the Indian people can speak and understand more than one Indian language as well as English. Hence even in the absence of OM to native font converters, people around the globe can type

and publish texts in the OM scheme that can be read and understood by many others, even when they cannot read native script. The readability criterion that has benefited from the case-insensitive phonetic mapping thus proves very useful. The major contribution of OM is to separate storing and rendering, which makes it language-independent across the Indian languages.

The OM mapping tables for many Indian languages can be seen at <http://www.dli.ernet.in/Om/>.

Om text editor

An integrated transliteration package that accepts OM ASCII keystrokes as input and maps them to native fonts has been developed. The script in any one of the supported true type fonts is sent to MS Winword® for further formatting and layout options. Since the OM scheme is common to all the Indian languages, the display of the text can be converted between the supported languages by choosing it on the menu. The tool also integrates with email clients on the windows platform. A web interface with similar functionality has also been developed. The text may be saved as OM (ASCII) text, native-font text or in Unicode. The tools have been used extensively for data entry for texts that feed into applications such as machine translation and optical character recognition. It has also been used purely for content creation by the outside community. An example may be seen at the magazine section of <http://www.telugumn.org>, where the story of Ramayanam has been created using this software. The integrated editor will be provided for hosting or use at any website free of cost, such as has been done at <http://www.telugumn.org>. The integrated editor is available for both windows and linux platforms. For those who wish to create content using a web interface, without the need to install the package, a java-based web interface is also available. The OM transliteration integrated editor is available for download at <http://swati.dli.ernet.in/om/>.

Multilingual book reader interface

The books on the digital library of India are available to *anyone, anytime, anywhere*. The goal of this work was to add to it the dimension *of any-language*. While OM transliteration helps one to read the text of one language with a script of another, it does not provide any translation. Due to the grammatical and etymological similarity amongst Indian languages, and their phonetic similarity, OM takes things a step beyond mere transliteration. The understanding can be improved further by the simple technique of merely translating some of the frequently occurring words from corpora. This has been the motivation behind the development of a multilingual book reader that supports *automatic transliteration* and *word to word translation* between Indian languages and between an Indian language and English. A Universal Dictionary, with the objective of providing the digital library user with a "good enough comprehension" of the content of the books in languages other than his own, has been built. This Universal Dictionary, which is cross-lingual, currently contains six Indian Languages (Hindi, Telugu, Assamese, Tamil, Kannada and Malyalam) besides most of the European languages.

When presented with electronic text in any Indian language, the book-reader allows text to be transliterated into any one of the many Indian languages. This is made possible with the OM transliteration scheme discussed above. This allows the user to read, for example, Hindi text in Telugu font. With the help of the Universal Dictionary, a word-to-word look-up table translation is made on the Indian language text between any pair of the many Indian languages supported by the Universal Dictionary. When a word is not found in look-up table, only its transliteration is displayed. While Indian

languages are phonetic languages, English is not phonetic. In order to display English words in native language where required, a pronunciation dictionary is used.

These features provided by the interface are desirable not only to the readers who can understand but not read their own language, but also to those who desire to obtain at least a crude translation of a book to their desired language. The book-reader performs the functions involving transliteration independently while also connecting to the example-based machine translation system on the backend for full-text translation. The reader, while especially suited to a multilingual country like India, is also extendable to any other digital library where the resources of translation and transliteration are available at large. The multilingual book-reader presents novel features that improve the usability and reach of any digital library.

Indian language Search Engine - Tamil Search Engine (OmSe)

Technology for the deployment of information retrieval in Indian languages has been demonstrated by the development of the OmSe search engine using the off-the-shelf open source software *Greenstone search engine*. The use of the OM transliteration scheme makes it possible to store any Indian language information in ASCII and to use any conventional search engine for information search and retrieval. For illustrative purposes, Tamil documents stored in the ASCII representation of OM have been built and are directly available for indexing, search and retrieval without any modifications to the text-handling modules of the search engine. At the time of display, the retrieved text, in addition being made available in readable English transliteration, is also converted to native Tamil script and displayed. Currently commercial quality optical recognition software is not yet available for Tamil and other Indian languages. Hence to demonstrate the technology, the Tamil search engine is built over a collection of born-digital newspaper articles, crawled from the web [3]. The basic architecture of the Tamil search engine includes a server that contains a database, a web crawler that crawls and downloads Tamil-language web content from various Tamil web portals, and the OM transliterator that converts true type font ASCII to OM transliteration format.

The front-end of the search engine is the client side, which has a graphical user interface that prompts the user to type in the search query in OM transliteration format. The query typed by the user is also displayed in Tamil font to enable the user to make corrections, if required, while entering the keyword in OM Transliteration format.

The interface between the client side and the server side consists of matching the user query with the entries in the database and retrieving the matched web pages to the user's machine. The search engine takes the query to the database and looks at the matches as per ranking. The search engine then sorts these database entries using a ranking algorithm. Greenstone's ranking algorithm determines the relevancy of a retrieved webpage to the user query. The retrieved sites are then displayed along with links to these sites in text format.

Machine Translation

Example Based Machine Translation (EBMT) is basically translation by analogy. An EBMT system requires a set of sentences in the source language and their corresponding translation in the target language. A bilingual dictionary comprising of Sentence Dictionary, Phrases Dictionary, Words

Dictionary and Phonetic Dictionary is used for the machine translation. Each of the above dictionaries contains parallel corpora of sentence, phrases and words, and phonetic mappings of words in their respective files. These files would form the database for the machine translation system. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to occur again. However, it is not possible to store the database as a set of a huge number of sentences. Instead, we can store the frequently occurring phrases and their translations and use these translations to translate a complex sentence. A sentence can be seen as a combination of phrases. Each sentence can be divided into a set of phrases and words. Instead of storing the entire sentence translation in the database, it would be more efficient to store phrase translations and word translations. This would optimize the database required for translations.

The EBMT has a set of 75,000 of the most commonly spoken sentences originally available in English. These sentences have been manually translated into three of the target Indian languages – namely Hindi, Kannada and Tamil. Bilingual word and phrase dictionaries between these target languages and English of over 25,000 and 18,000 entries. respectively, were also created manually. The artificial intelligence engine learns from these examples and provides a good enough translation by looking for the longest match at the sentence, phrase and word levels. An even better way of storing the database is to store the rules related to the language pair. This has been observed to improve results dramatically. Though we would say that it is EBMT, the rules database would drastically reduce the size of the database and would actually improve the translation results. Thus the phrase translations and phrase rules play a significant role in machine translation. The advantage of this simple "good enough translation" system is that its performance can be improved almost linearly with the increasing corpus and rule base.

The web enabled version of the EBMT is available at <http://bharani.dli.ernet.in/ebmt/>. The current machine translation system supports the following language-pair translations:

1. English to Hindi
2. English to Kannada
3. English to Tamil
4. Kannada to Tamil

The EBMT also has an interface for learning rules from its users. User feedback is stored separately and used later, after a verification for correctness, so that the main data base, and hence the performance of the machine translation system, is constantly improving.

Conclusion

The Digital Library of India (DLI), besides providing an opportunity to create freely browsable and searchable documents of value to all of humanity, has become a testbed for Indian language research. DLI has developed search engines, and clusters for data base storage and retrieval. It has also helped in creating a tight bond for those doing research in India within 21 centres, spanning academia, government and religious institutions. It is heartening to see that in India the Government, policy makers, religious institutions and scientists have put India on the world map of digital libraries and have done so as a national project under the Ministry of Information and Communication Technology. Many organizations are waiting to join this national initiative.

Acknowledgments

Several students and staff at IISc, CMU and IIITH, and from the participating centres, have contributed to every aspect of this national initiative. Funding for the Digital Library of India comes from multiple sources. The Office of the Principal Scientific Advisor to the Government of India is funding the project at the Indian Institute of Science. The Ministry of Communication and Information Technology (MCIT) is funding the project at various partner centres of the Digital Library of India. Various centres have also pledged their local resources to make the Digital Library of India a reality. The National Science Foundation (USA) is providing funding for scanners and software Research and Development through Carnegie Mellon University. It is also fortunate that the First Citizen of India, His Excellency Dr APJ Abdul Kalam, who himself is one of the contributors to this vision, has personally taken an interest in making the Rashtrapathi Bhavan one of the major centres of the Digital Library.

References

1. Ingo Fromholz, Predrag Knezevic, et al., Supporting Information Access in Next Generation Digital Library Architectures. In *Proceedings of the Sixth Thematic Workshop of the EU Network of Excellence DELOS(2004)*.

2. Rothenberg, J. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Rep. to Council on Library and Information Resources, January 1999.

3. Anandh Jayaraman, Srinivas Sangani, Madhavi Ganapathiraju and N. Balakrishnan, OmSE: Tamil Search Engine, In *Proceedings Tamil Internet conference*, pp 23-29, December 2004.