Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

Universal Digital Library—Future research directions

BALAKRISHNAN N.

(*Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560 012, India*)

E-mail: balki@dli.ernet.in

Received Sept. 30, 2005; revision accepted Oct. 3, 2005

Abstract: This paper starts with a description of the present status of the Digital Library of India Initiative. As part of this initiative large corpus of scanned text is available in many Indian languages and has stimulated a vast amount of research in Indian language technology briefly described in this paper. Other than the Digital Library of India Initiative which is part of the Million Books to the Web Project initiated by Prof Raj Reddy of Carnegie Mellon University, there are a few more initiatives in India towards taking the heritage of the country to the Web. This paper presents the future directions for the Digital Library of India Initiative both in terms of growing collection and the technical challenges in managing such large collection poses.

Key words: Digital Library, Search engines, Language technology
doi:10.1631/jzus.2005.A1204 **Document code:** A **CLC number:** TP391

INTRODUCTION

The stunning growth in storage technology prompted Prof. Raj Reddy of Carnegie Mellon University to envision that it would be possible to store in digital form, in the very near future, all knowledge ever acquired by the human race. As part of this grandiose vision and as a first step, in realizing this vision, it was proposed to create the Digital Library with a free-to-read, searchable collection of one million books, predominantly in Indian languages, available to everyone over the Internet. In this worldwide mission, USA, India and China are some of the major contributors. The Indian effort is named as the Digital Library of India. The Digital Library of India (DLI) is expected to foster creativity and free access to all human knowledge. This portal is also planned to become an integrator of all the knowledge and digital contents created by other digital library initiatives in India and in other partner countries such as China. This portal (http://dli.ernet.in and http://dli.iiit.ac.in) is slowly and steadily becoming a gateway to Indian Digital Libraries in science, arts, culture, music, movies, traditional medicine, palm leaves and many more. Currently more than 120 000 books (around 50 million pages) have been scanned and most of them are available on the Web for free browsing.

One of the goals of the Digital Library of India is to provide support for full text indexing and searching based on OCR (optical character recognition) technologies where available. The availability of online search allows users to locate relevant information quickly and reliably and possibly in a language independent way, thus enhancing student's success in their research endeavors. In order to achieve this mission, CMU and the Indian Institute of Science along with 21 partner institutions, first established technologies and processes for the selection of books, their scanning and cropping, OCRing and the storage architectures. Besides acting as a repository of information, the Digital Library of India had also become one of the finest test-beds for Indian language processing research in areas such as machine translation, optical character recognition, summarization, speech and handwriting recognition, intelligent indexing, and information retrieval in Indian languages.

TECHNOLOGICAL CHALLENGES

The DLI when completed, will host enormous amount of quality data in Indian languages. This is proposed to be used as a test-bed for Indian Language Technology research. The large volume of data available today has already stimulated many new and innovative initiatives in India. Major technologies that have been developed so far include:

(1) The Om Indian Language transliteration package which for the first time introduced a case insensitive scheme of representation of Indian language contents and also exploited the phonetic nature of Indian Languages to separate the storage and rendering. The storage is independent of the language while during the rendering, a choice of Indian language can be made by the user of Om.

(2) The Universal Indian Language book reader, which starts of with a simple transliteration. This helps the users to read documents in other languages and increases comprehension using the similarity between many Indian languages. The useful extensions to this include augmentation with the universal dictionary and an Example based good enough and simple to construct machine translator.

(3) An example based multi lingual machine translator.

(4) The Indian Language Search Engine, which retrieves documents in any Indian language using a public domain software − Greenstone.

(5) Some other issues connected with speech synthesizer in Indian languages are also currently receiving attention at the Indian Institute of Science.

FUTURE DIRECTIONS

In the first phase of the work the DLI has concentrated mainly on books and it has almost started to look scanner centric rather than knowledge centric. But with the inclusion of many of the religious institutions as partners in the DLI, other forms of traditional knowledge information are also getting added to the collection. These pose many newer challenges, which are likely to be more unique to the Indian heritage.

Much of the old information of heritage value has been written on palm leaves and metal plates. They also use alphabet and words that are no longer in use. The technology for scanning these is more complex than those needed for printed text and also handling of centuries old artifacts call for special attention. It is almost impossible to design OCRs for palm leaves. Currently, in order to circumvent the complex design of OCR, a novel approach of first using humans to read the documents and then using a speech recognizer to index the speech data is being pursued with very encouraging results.

The knowledge base and the content creation in India have also encompassed the codifying of ancient or traditional knowledge. One of the most successful efforts in India is that of the CSIR in India and it is called the Traditional Knowledge Digital Library (TKDL). There is also an initiative sponsored by the Government of India to provide a common framework for all the Digital Libraries in India to coexist.

Indian life revolves around sports and entertainment and India is one of the largest producers of movies in the world that too in many languages. In tune with the exponentially growing storage and bandwidth, it is proposed by Prof Raj Reddy to digitize music, video and other sports, entertainment and religious discourses both live and archived. This will enhance the utility of the DLI and also expand the definition of library to become truly a powerhouse of information. This will help also in documenting many traditions and localized information that in the Indian context has been going on from generation to generation by word of mouth, folklore and songs.

The inroads into entertainment and music will also bring in many commercial challenges that have to be met. The DLI is working under the guidance of Prof Raj Reddy on a novel idea of a “Consortium for Compensating for Crating Contents” – the FourCs. This scheme addresses the possibility of creating a 21st century equivalent of Public Libraries, the PBS and the All India radio to create contents for the Web. This will also feed critical inputs upon which the copyright issues for scholarly communication, historical and heritage information systems.