# Indian Digital Library Movement: Need of Indian Initiative

Aditya Tripathi

**Abstract**

*The technological advancements have enabled people of India to be heard over Internet. The country has a large treasure of written text which warrants a wide circulation and communication. But there are many inherent problems appear at the time of mechanical text processing and representation of Indic scripts. The paper highlights the inherent problems of Indic text processing and suggests a nationwide movement to address them. Paper presents an investigative account of difficulties towards creating digital libraries in Indian languages.*

**Keywords :** Digital Library, Indian Languages.

## 1. Introduction

India has a glorious history of 5000 years with a rich heritage of philosophy, language, religion, art and culture. The Nation has 428 languages with 1600 dialects out of which 13 are extinct [1]. These languages have several classical scriptures in their Scripts. These scriptures are scattered and preserved in different libraries of the country. Creating a digital library of available classics and establishing federated look-ups to these collections are burning issues of the day. The Indian Digital Library Initiative (IDLI) has been triggered out of Digital Library Initiative (DLI) of United States. Both the Indian collections and the environment are totally different from the one United States has. A fresh look to the existing problems should be given.

## 2. Issues with Indian Digital Libraries

Though there is a huge thrust in India for establishing digital libraries, there are several reasons due to which the whole movement is fallen behind.

### 2.1 Behavioral Issues

The Government of India must look forward to pool the knowledge base of the country. Phase-wise pooling of available knowledge resource of various libraries in the country must be encouraged. This requires funding to various libraries for digitizing the documents of their collection. So far, much of the funding for the purpose is made through United States through Million Book Project. Though, the project has been initiated with novel interest, unfortunately, keeping control over the whole project appeared little difficult. Besides, many of the technological challenges specific to Indian context are new to the American environment need to be addressed.

It looks better to initiate test-beds in first phase with the objective to demonstrate technological capabilities. Some centres can be identified from different language speaking states and accordingly funded for developing technical capabilities in a limited time frame. Unfortunately, the present

 

growth of digital libraries in the country is unplanned which has resulted in bad models or instances with just scanned images. The first phase of the National Digital Library Movement could have been better utilized only to understand the underlying difficulties and the possible solutions.

The second phase of the project could have been funded for building the live digital collections for use and receive feedbacks. The required changes and modifications could be further implemented nationwide.

There are several libraries with unique collection of its kind. With the third phase such collection can be digitally preserved as well as universal access to the documents could be insured.

It has been observed that many of the libraries having unique and rare collection are scared of going online because they think of loosing the physical presence of users in the library. This behavioral difficulty must be addressed in order to ensure universal accessibility including the digital preservation at national level.

## 2.2 Conditional Issues

A multilingual country like India has several conditional constraints hindering the overall growth of digital libraries in the country. There are representation in the literature from all states and languages. It is easy to map the script due to common origin from Brahmi but it is equally difficult to convey the meaning in different languages. It can be easily observed that in different parts of the country people know more than one language with having one script. For example, it is very common to see Brahmins reciting hymn in Sanskrit (a language) written in their mother script. In such cases mapping of characters can be very handy.

**Fig 1:** A page in Sarada script from Kathaka-Grahyasutra

(Courtesy Sayaji Rao Gaekawad Library, Banaras Hindu University)

Since writing is known to Indians from 3000 BC and paper is invented only in 104 AD. Apart from that the printing technology reached very late to India. Hence, it was very late when Indians started using paper. Many of the ancient writings are on stones, leaves, clay tablets, tree barks, cloth etc. Anything before, 1600 AD is hand written which is difficult for machines to read. Some of the objects are so brittle that they can only be handled using special means. There is abundance of such material in the country.

## 2.3 Technological Issues

World has witnessed impeccable technological advancements. There are digital library suites can be used directly for creating digital collection. But much of the development is in English and European languages as the computer was first used by them only. But off late, market for other scripts is also felt by IT giants and hence Unicode as character representation standard is introduced. Unicode is still being evolving as it promises to cover all the world scripts and signs.

The country has a big share in IT industry of the world. But little has been done for the languages and scripts of India. An Optical Character Recognitions (OCR) software in any Indian script is yet to be developed. The tests are on the way at Indian Statistical Institute, Kolkata and Centre for Development of Advance Computing (CDAC), Pune. A product named Chitrankan by CDAC, Pune is available for use in Devnagari but the product has limited variety of font support and is still in testing phase.[2]

The Indic scripts are off child of Brahmi script which is not in practice anymore. The derived scripts of Brahmi like Devanagari and other Indic scripts have complex rendering principles when it comes to representation of characters in machine. Complex character glyphs are rendered depicting features like bidirectional and dynamic composition.


**Fig 2:** Bidirectionality and Dynamic composition in Devanagari script

It is difficult to create rendering engines of this kind for all the Indic languages but major work has been done in this area. The work is in progress to integrate rendering plug-ins with web browsers for creating websites in Indian scripts. With Devanagari and other major scripts the task has been achieved and now-a-days one can see many websites in regional languages.

One of the important problems Indian IT community facing today is about sorting the words of one language. The order given by Unicode does not suit with the order of sorting in practice. For example, character क्ष comes in the end according to practical sorting where as in Unicode it comes after character क, because character क्ष is generated by the combination of three characters i.e. क + ् + ष . There are scores of examples of same kind in Sanskrit. It is because of these

reasons it is very difficult to generate index for any Indic document. Besides, there are no good corpora available in public domain which can be taken as sample for experimentation and testing. This has also influenced the search algorithm.

Unlike search algorithms of English, search algorithms in Indian languages are yet to be evolved. What is available today as Indian language search is only pattern matching. The work is still at preliminary level because of the complexity of the languages. It is easy to work with Sanskrit for the purpose as it has morphological rules for construction of words but addressing the same is difficult with other Indic languages (Hindi, Oriya, Tamil etc.) as they don't have such kind of rules for word construction. The Stemming algorithms like Soundex and Metaphone are yet to be developed by any of the Indian languages. To get a feel, if one searches for Hindi word पिछे (correctly spelt as पीछे, means Back) in Google, gets only the former Hindi string ( पिछे ) in search result. An application of stemming algorithm would have brought the correct later one (पीछे) in the search results. Similarly, there are n-gram techniques of string search which are warranted to be tested with Indic text.

Though the metadata is not a potential problem but due to uncontrolled growth there is lot of disparity in use of metadata standards. Many of the libraries have drafted their own metadata set not confirming to the International standard. This has happened in the case of distributed digitization centers at the initial stage. Though, an awareness has been aroused among the professionals to use International standard, but, still it is very difficult to convert nonconforming one's to the standard form. With the use of codified data a standard metadata scheme can yield an efficient cross-lingual information retrieval. This technique of cross-lingual information retrieval has been demonstrated in the system named "Brass" developed at Documentation Research & Training Centre [2]. Such information retrieval systems can supplement automatic translation projects to some extent, as there is not significant development in machine translation of Indic languages.

## 3. Indian Scenario

India is witnessing a strong move for developing digital libraries around the country. Some are in public domain like LDL of Documentation Research & Training Centre, Bangalore (see at https://drtc.isibang.ac.in) and Digital Library of India (see at http://dli.iiit.ac.in/). Some are with restricted access like, Nalanda Digital Library of National Institute of Technology, Calicut (see at http://

www.nalanda.nitc.ac.in/index.html). As far as, the English literature is concerned there are editable text available over web but when it comes to Indian language only scanned images of the documents are available for display supplemented by metadata.

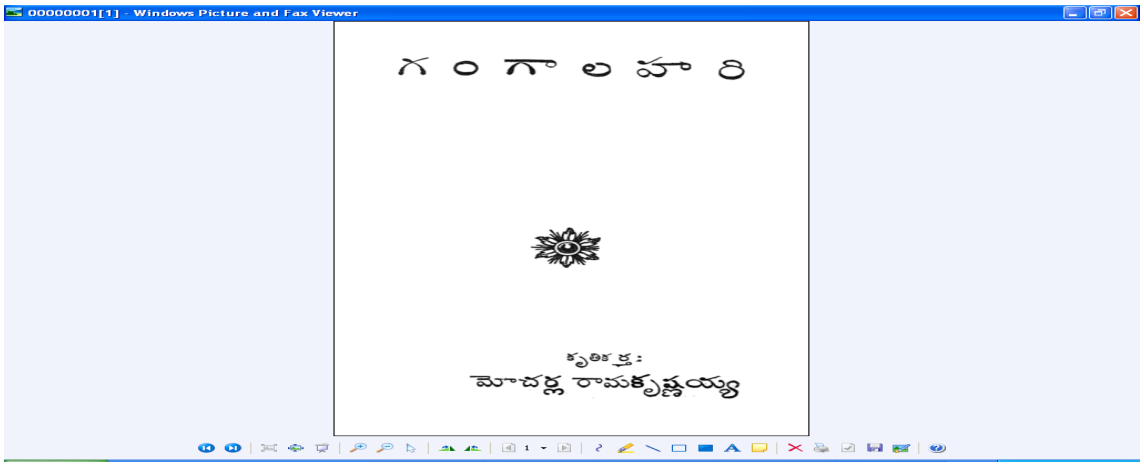


**Fig. 3:** A scanned page of a Telugu document from Digital library of India

The experiment has been done with many of the Open source software and results are positive. Some test beds can be found at Documentation Research & Training Centre site but still much has to be done.

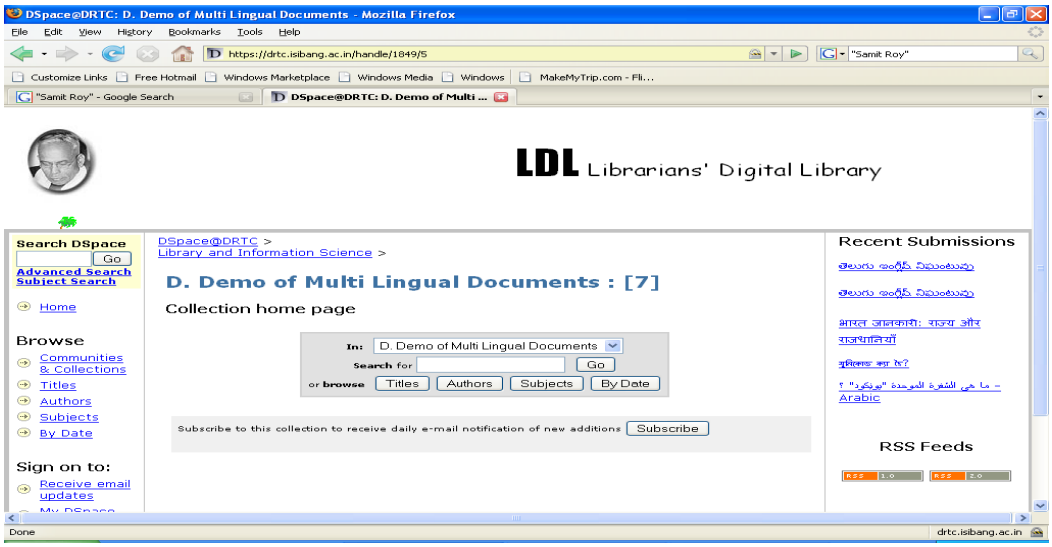


**Fig. 4:** Test-bed of Multilingual Digital Library at DRTC

## 5. Conclusion

With the advent of multilingual character representation in the computers and its harmonization with ASCII (American Standard Code for Information Interchange) has supported vast amount of multilingual data to be transferred over Internet without producing network congestion. Libraries can well utilize the present scenario and publicize their online document delivery in regional languages.

The overview of Indian environment brings up many potential area of research funded by the Indian government through National Digital Library Initiative such as metadata standardization, developing and testing search algorithms in Indian languages etc. A commendable encouragement is given by Department of Information Technology under Ministry of Communication and Technology. INFLIBNET can be a pioneer in establishing a policy towards supporting university libraries for enhancing multilingual communication. Hope Indian community would see some remarkable developments in near future.

## References

1. Languages of India. Available at http://www.ethnologue.com/ show_country.asp?name=IN (accessed on 14/01/2008)
2. Tripathi, A. Design and Development of Multilingual Information Retrieval System with Numeric MARC. Doctoral Desertions. 2004
3. UNICODE CONSORTIUM. The Unicode 5.0 standard. Available at http://www.unicode.org
4. Nurturing the Living Languages. Available at http://www.cdac.in/html/gist/gistidx.asp (accessed on 14/01/2008)
5. Chitrankan. Available at http://www.cdac.in/html/gist/products/chitra.asp (accessed on 14/01/2008)
6. Google. Available at http://www.google.com

## About Author

**Dr. Aditya Tripathi,** Lecturer, Department of Library and Information Science, Banaras Hindu University, Varanasi.
E-mail : aditya@bhu.ac.in; adityatripathi@hotmail.com.