

Nitya Archive: Software for Full Text Digital Libraries in Indian Languages

R. Raman Nair and K H. Hussain

University Librarian
Mahatma Gandhi University
Priyadarsini Hills
Kottayam

ABSTRACT

The paper looks into the need for developing full text digital archives and libraries with programmes that can process Indian vernacular languages. Use of local scripts becomes inevitable in the creation of Information systems for digitized palm leaves, manuscripts and other local documents. Also discusses about the importance of conceiving an alternate algorithm for digital archives that is capable of searching and retrieving individual pages rather than whole documents, which is the current fashion among international packages like GreenStone, DSpace, EPrint, etc. Traces out the historical development of the package 'Nitya Archives' which can be used to create multi lingual digital archives in Indian languages and describes in detail the major projects implemented using it.

Introduction

During the last two decades many research and academic organizations created electronic catalogues of books in local languages using English transliteration. Databases were created using free CDS\ISIS or other commercial packages. But databases for documents in local scripts were not efficient due to the default English script of computer. Since most of the projects did not input data conforming to a standard transliteration scheme, words and phrases formed at the time of query differed from they were coded at the time of data entry. This led to non-retrieval of titles in the collection. Even though some of the libraries devised transliteration schemes, data in different Indian languages became so un-natural that even the library professionals could not decipher them. Developing Information Systems in Indian vernacular languages is now critically felt more than ever. Use of local scripts become inevitable especially in the creation of Information systems for digitized palm leaves, manuscripts and local documents.

Language Technology

GIST technology devised by C-DAC in the mid 80's was a commendable attempt that circumvented the above-mentioned situation. Indian language implementation by GIST was a hardware solution, which helped to create a lot of databases in local languages using dBase2. Later this technology was replaced entirely by software based on 'Key board hooking', an interface that enabled the data input in any running application like word processors, DTP packages, DBMS, etc using any Indian script. In many databases data could be entered in local scripts. But in the search module the interface often failed and the index appeared in unrecognizable Roman characters. So search became an unattainable goal after the 'advanced' GIST technology.

Real solution lies in embedding Indian scripts in operating systems like Windows, Linux, etc. Microsoft could embed most of the scripts of Indian languages by 2003, thanks to the new Unicode encoding. It was Unicode encoding and OTF (Open Type Font) technology that finally set the multilingual problem in the right track, which could not be resolved by ASCII encoding.

During the last five years various NGO's from different parts of the country have been successful in designing Open Type Fonts (OTF) for different Indian languages. Linux presents a more open, strong and net-workable environment compared to Microsoft's. But it will take a few years to popularize Linux and to have full text archive management packages.

Full Text Archives Development

Many initiatives have come in India for Full Text Digital Libraries since 1995. Even if we consider the case of full text in English they all fall short of expectations. So developments on this line for full text in Indian languages were very few. Some of the important initiatives are listed below.

The first full text digital library was envisaged in India in 1995 under Dr. A. M Michael, Vice Chancellor of Kerala Agricultural University (KAU). It was the digitization of PhD theses approved by KAU and the project was a subset of the Kerala Agricultural University Library and Information System (KAULIS). With support from ARIS programme under ICAR, KAU digitized 400 of its 3000 dissertations in 1997-98. At that time even the term digital archive was not heard in India. Unix based TechLib Plus was the only package found to launch digital collection in LAN and for a short period KAU archive was available in the Intranet and Internet. The Vice Chancellor and the staff who initiated the project left the organization due to political interventions. Even though ICAR was prepared to provide funds for support and development of the digitization project, remaining dissertations were not added to the archives. In 2000 one of the earliest digital library ventures in India thus went non-functional along with KAULIS project for which an amount of 9.5 crore rupees had already been spent for infrastructure facilities.

Role of CDS/ISIS

Though applications of CDS\ISIS in Indian libraries are declining, its superiority as a textual database and documentation package is undeniable. 'Nitya', a package programmed using ISIS32.DLL for digital archiving have already illustrated the strength of CDS\ISIS in retrieving full text. 'Nitya' in its early years explored potential of CDS\ISIS in CD-Publishing of databases and full texts. Adapting to Indian languages the UNESCO documentation package can be better utilized. M-ISIS, the Malayalam version of CDS/ISIS developed by Nitya team was an attempt in this direction.

Brennen CD

The first bibliographic information system using M-ISIS was created in the library of Government Brennen College, Thalassery, Kerala. Malayalam collection in the library, one of the oldest in Kerala, numbers to 21000. Its catalogue using Malayalam script (ASCII Character set specially designed) was created using M-ISIS and whole database was published as a single CD in 2004. This bibliographic CD is the first of its kind in Kerala and the content can be searched by Author, Title, Subjects, etc using Malayalam Script. Brennen-CD showed the feasibility of building up information systems of Malayalam documents using Malayalam script. The use of Original Malayalam script advocated by 'Rachana Akshara Vedi' showed that it is the most standardized and comprehensive character set, and hence advisable for creating information systems in Malayalam. After achieving search and retrieval using Malayalam script, next attempt was to combine the technique of M-ISIS and 'Nitya' to create a digital archiving system.

M-ISIS and 'Nitya'

Though the dissertation archive under KAULIS was not sustained, this project infused interest in many organizations in ICT field. Under an NGO named Centre for Informatics Research and Development (CIRD) further research on digital archiving progressed. CIRD developed a powerful package in 2005 named 'Nitya' that could process Malayalam and English scripts for full text digital archiving.

M-ISIS was a localized version of CDS\ISIS for Malayalam language while 'Nitya' was a search and retrieval package able to open the full text in different format like, PDF, JPG, DOC, TXT, etc. M-ISIS was a reference retrieval system where as 'Nitya' was a full text retrieval system. Both used CDS\ISIS as their database engine. The front end was programmed in Delphi (Object Pascal) using ISIS32.DLL created by UNESCO and BIREME (Latin American and Caribbean Center for Health Science Information).

KAU Theses Library in Nitya

As a first experiment the digitized dissertations at KAU had been exported to the earlier version of Nitya by CIRD in 2000 voluntarily, which revealed the power and possibilities of the emerging concept of digital archiving. Later the package Nitya was improved in various digital archiving contexts. By then various free digital archiving software like GenISIS, Greenstone and DSpace have come into the scene. But Nitya remained unique for Indian contexts due to its power to process local scripts.

A sample Reference library on Kerala was also developed exporting the Database from Brennen CD and related original documents digitized and saved in PDF format. With the success of this experiment Centre for South Indian Studies (CSIS, Thiruvananthapuram) has initiated with technical support from CIRD a digital library project entitled 'Kerala Reference Library' consisting of selected authoritative rare and antique books on Kerala's history and culture in Malayalam and English. The project when completed will make an authentic and reliable reference collection on Kerala affordable to more than 6000 rural libraries as well as school and college libraries in Kerala.

Mathrubhoomi Weekly Archives

Next project attempted with Nitya was the development of an archive of back volumes of 'Mathrubhoomi' weekly published since 1923, which is one of the important literary and cultural magazines in Malayalam. A few issues of the weekly were digitized and saved in PDF. Catalogue data of every individual article was fed into CDS\ISIS database through a data entry worksheet designed in M-ISIS. 'Nitya' performs search and retrieval of the full text. Later the whole database was recreated using Unicode Malayalam replacing ASCII font of M-ISIS. This prototype of Mathrubhoomi Archive can be used by Mathrubhoomi or similar Malayalam journals to archive their whole collection.

Unicode Age 2003

By this time Unicode standards have come into existence, which made possible of processing main Indian scripts in computer. State Central Library (Trivandrum Public Library) used the Unicode compliant 'Nitya' for preparing an in-depth catalogue of 800 old rare books in Malayalam and English for digital archiving the collection. It is the first Unicode based Multilingual DBMS developed in Kerala. It used Original/Old Malayalam script applying 'Rachana' font.

Trivandrum Public Library

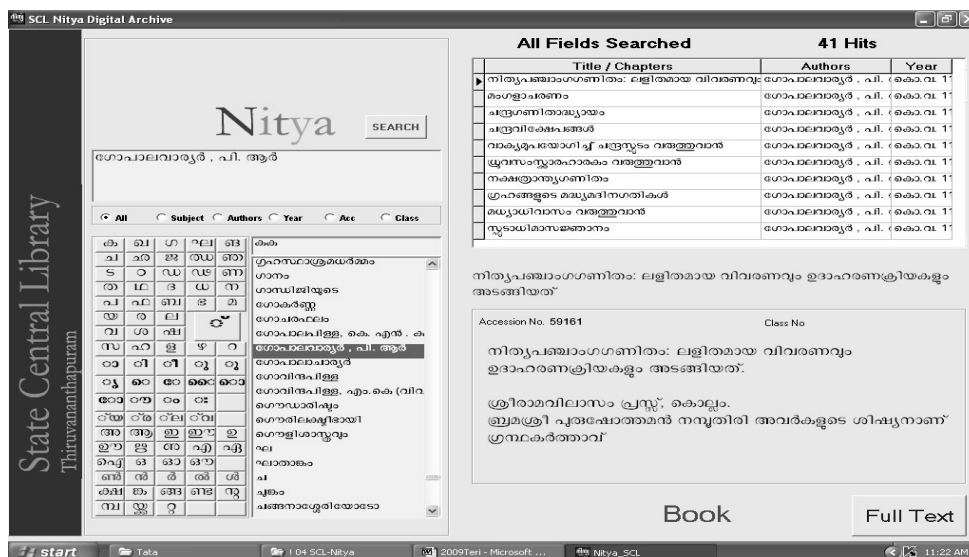
Established by Maharaja Swathi Thirunal in 1829 Trivndrum Public Library is one of the oldest Libraries in India. The library is having a substantial collection of the rare books not available in any libraries in India. More important than that many of these rare items are of high relevance to research on the region. The books in different languages like, Tamil, Malayalam, Kannada, Hindi and English. In 2004 Government of Kerala decided to digitize the rare collection and Nitya was selected because of its multilingual capacity for handle local Indian scripts effectively.



Home Page of Trivandrum Public Library Digital Library

Nitya customized for Trivandrum Public Library Rare Collection with user-friendly front-end can search and retrieve using English as well as Malayalam, Hindi and Tamil scripts.

In the search page user can first select the category like author, title etc. In the search screen user gets an inverted file in concerned language and can click the word to form query or directly type term in local scripts using online keyboard.



Search mechanism of TPL Digital Library

Clicking the search button retrieves a list of titles available in the collection. Clicking each title yields full bibliographical information in concerned language. Full text can be opened in Acrobat reader by another button click.

Digitization Project Report for Kerala Legislative Assembly Proceedings

Legislative Records mainly consist of the deliberations of Legislative Councils / Assemblies in the form of proceedings and various Committee Reports. Proceedings contains various sessions like Question-Answers, Adjournment Motion, Calling Attention, Submission, Financial and Legislative business, Points of Order, Rulings, etc. Nature of their content, periodicity of their issue, their importance as authoritative sources for legislative process, etc. are unique and they form one of the most important official record collections held by government.

Earlier documents in Kerala Legislature are mainly hand written or typewritten. A few of them have only single copy and often in brittle condition. Any defect caused to these documents will result in the loss of unique items related to Legislation. They are constantly in demand and are subjected to continuous Xeroxing. Frequent handling of proceedings will gradually lead to the destruction of these documents.

Retrieving relevant information from these unconventional publications is a Herculean task. Over the years, the legislative documents have accumulated in thousands of volumes without proper organization, bibliographic control and content indexing.

In the day-to-day functioning of the Legislature these documents are heavily consulted by Speaker, Ministers, Legislators and officials. As the quantum of Legislative documents is growing, retrieving information from these documents is becoming more and more complex. Unless elaborate and in-depth indexes are prepared majority of the content will go unnoticed.

Presently nearly 6.2 lakh printed pages of Kerala Legislative Assembly Proceedings are available. Nearly 75000 pages are in English, Tamil and Kannada. Of the documents 75% are in old original Malayalam Script. So the digitization of this collection requires a multilingual program that should handle English and three local scripts together with a system of deep context indexing.

In response to a request from Kerala Legislature in 2007 Centre for Informatics Research and Development (CIRD) consisting a team of specialists on digital library and language technology analyzed the existing system and a detailed technical project report was prepared. A survey of the digital library systems in Kerala having multilingual capability was conducted and a report with full technical details and specification was prepared. The archive was developed in accordance with these specifications.

Further customization of Nitya was essential in the case of Legislature Assembly proceedings. A particular day's proceedings bears diverse access points like business transacted, bills introduced, papers laid on the table, reports and budget presented, motions moved, special discussions held, allegations raised, rulings given, names of members participated, subject of discussion, etc. All these aspects should be distinctly structured into a database without which the digital files of proceedings would be useless. The current package programmed in line with Nitya by the recommendations of the project report materializes above objectives.

Mahatma Gandhi University Digital Archives

Established in 1983, MG University has awarded about 1300 doctoral degrees in various disciplines during the past 25 years. In its Silver Jubilee Year 2008 University initiated the development of an Open Access Digital Archives of PhD Dissertations approved by the university. It is the first official, authentic and reliable digital archive of doctoral dissertations of any Indian university. It hosts all the accepted dissertations available in the university. The project claims that all future dissertations will be uploaded within one week after their acceptance and issue of degree by the university.

Software Features

MGU has used the latest online version of Nitya Archive. Further customization of the package was done for this specific project by Beehive Digital Concepts Pvt. Ltd, Cochin. MGU Online Digital Archives of Dissertations is web hosted using this special version of Nitya, which has got UNICODE compliant multilingual search facility.

Specificity in search and retrieval offered is commendable and its metadata can be made OAI-PMH (Open Access initiative – Protocol for Metadata Harvesting) compliant. Following are the important features of the archive.

- The Archive can host any number of dissertations in English, Malayalam, Hindi, Sanskrit, Tamil and Kannada.
- Provides facility for multi keyword Boolean search
- Facilitates multilingual data input and query building
- Dissertations are not password protected to access and view
- No need to download the Dissertations to read them.
- Compatible with any browser on any operating system
- No need for additional software like java or flash. Only browser is needed.
- Accessible from anywhere at any time through the URL <http://www.mgutheses.org>

Technical Data of MGU Archive

- The Archives is hosted on a live and dedicated Linux Server, catering thousands of simultaneous requests without performance loss. The server is secure and available all time.
- The project is built on PHP programming language in combination with MySQL.

Unique Features

- Exhaustive collection of doctoral dissertations.
- Retrieval of specific pages/sections.
- Open database structure allowing migration to OAI-PMH compliant packages
- Query building and searching using English, Malayalam, Hindi, Tamil and Kannada scripts.



Home page of MGU Dissertation Archives at <http://www.mgutheses.org>

Using MGU Archive

Compared to the dissertation archives hosted in the net by institutions worldwide using DSpace, Greenstone and other packages MGU Archive has a simple and user-friendly interface similar to Google. Other packages retrieve the needed dissertations in one piece while searching for a specific keyword and users have to manually navigate the entire document. In MGU archive the search directly takes user to the concerned page of a dissertation and there after they can navigate effortlessly to other sections using bookmarks. For a simple search subject keywords are entered in the search box in the Home page. It can be words that may occur in the title or bookmarks in the dissertation.

Getting List of Relevant Dissertations



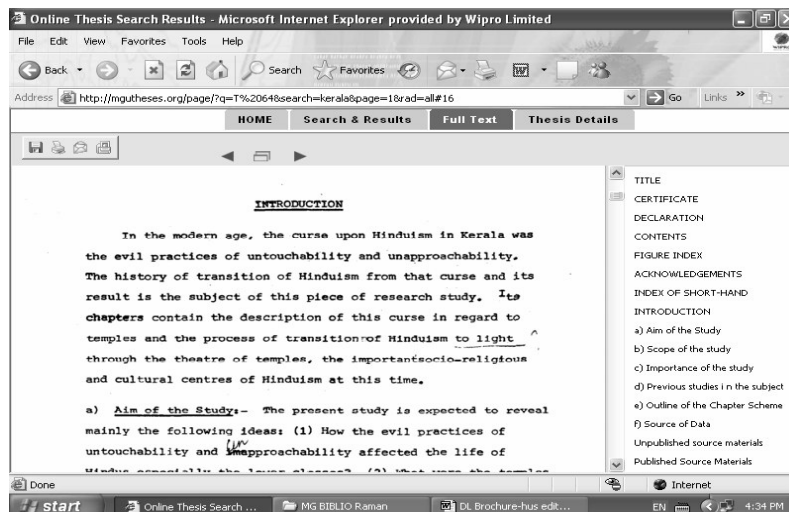
Clicking the 'Search' button, all the titles of the retrieved dissertations are first displayed. After that all the sections in other dissertations related to the query are listed.

The hit list of dissertations by a query can be printed for future reference. From the hit list user can click the required title and go to a particular section or entire thesis.

Navigating Through Full Text

Once the full text is opened user can navigate through entire thesis back and forth. Bookmarks provided on the right side of the screen helps to go to the relevant sections/chapters instantly.

Downloading the dissertation is not permitted presently. Subject to detailed discussion at MGU University and after watching the development that may occur in other Indian universities these permissions will be provided.



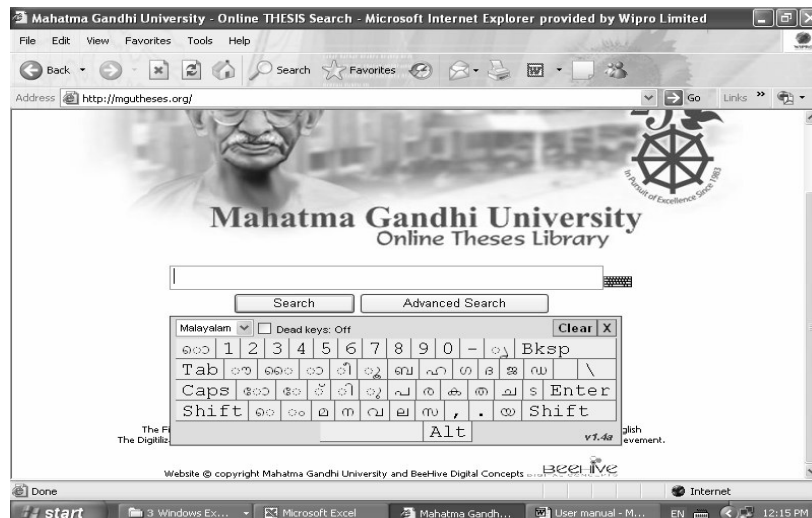
Advanced Search

Searches can be made selecting categories like Title, Scholar or Guide. 'Advanced Search' option lets the user to build complex Boolean queries.



Multi Lingual Search

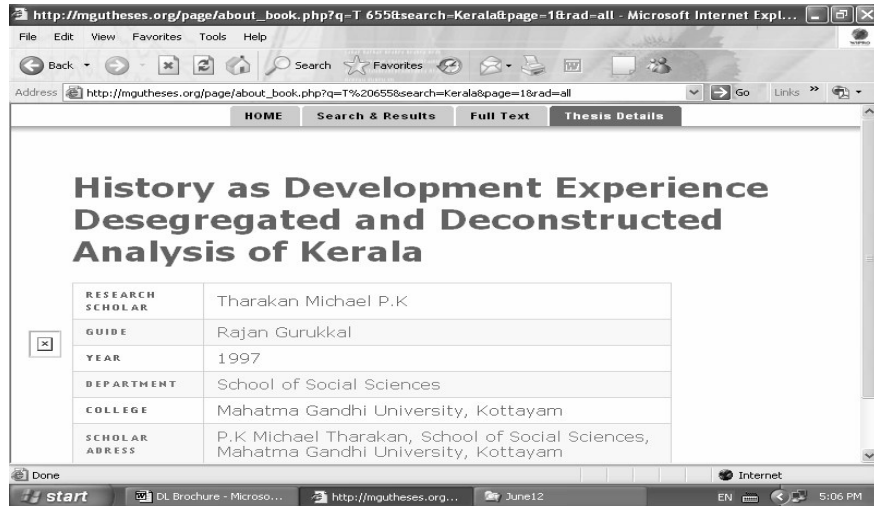
Since Online Nitya is Unicode compliant dissertations in any Indian language can be archived and retrieved using local scripts. MGU Online Dissertations Archives is the only bibliographic information system presently existing in Kerala having multilingual search capability using English, Malayalam and Hindi scripts. A visual Malayalam keyboard is provided to construct queries for those unfamiliar with Inscript keying.



Online Nitya will be an effective model for a regional dissertation archive for universities in India. Kerala University and Kannur University can now think of an effective way of archiving their Tamil and Kannada dissertations in line with MGU archive along with their English, Hindi, Sanskrit and Malayalam titles.

Bibliographical Data of Dissertations

Clicking the tab 'Theses Details', users get full bibliographical data of the thesis.



International Recognition for Nitya

INTUTE is a UK Based consortium of universities working with a whole host of partners, through which academic and research resources in the web are evaluated for inclusion in its system. The mission of INTUTE is to advance research by promoting the best of the Web through a process of evaluation and collaboration. INTUTE is funded by Government of UK's Joint Information Systems Committee (JISC), and the Arts and Humanities Research Council (AHRC). After evaluating the content and software mgutheses.org has been selected by INTUTE as one of the Very Best Web resources for education and research and included in <http://www.intute.ac.uk>. The selection information occurs at the INTUTE website. MGU is the only university from India that hosted multilingual digital archive of full text in the web using search mechanism approved by INTUTE.

Conclusion

Present retrieval model of 'Nitya' was conceived from the inability of other archiving programs to retrieve specific pages. Nitya has been evolved from its humble beginning in 2000 to an internationally recognized archival package. Nitya is a 'procedure' rather than a program, which has got three distinct steps. First, the documents are scanned as Acrobat PDF files and they are graphically corrected / modified. Second, PDF files are 'Bookmarked' with chapters, sections, tables, figures, charts, etc. and documents are organised matching to its hard copy. Third, metadata (bibliographic data) together with its bookmarks are taken to a relational database. The front-end 'Nitya' uses the 'bookmark table' to search and retrieve specific pages.

'Nitya' model for the retrieval of specific pages is still a unique method in digital archiving. Internationally acclaimed packages like DSpace, GreenStone, EPrint, etc have not yet attempted this concept. The popularity of the MGU Theses Archive hosted in Nitya among the scholars, students and teachers as well as the recognition it brought to the university from national and international organizations amply testifies the potential of this novel concept of deep indexing in digital archiving.

References

- Hussain K H, Vijayakumaran Nair P, Chitrajakumar R, Ravindran Asari K, Raman Nair R. Creation of Digital Archives in Indian Languages Using CDS/ISIS: Development of M-ISIS (Malayalam ISIS) and 'Nitya'. *Information Studies* 2005, 11(1), 59-68.
- Jancy James, Raman Nair R and Sreekumar G. Doctoral Research at Mahatma Gandhi University: A Bibliometric Analysis. *University News*, December 22-28, 2009.
- Mahatma Gandhi Open Access digital Archives of Doctoral Dissertations. <http://www.mgutheses.org>
- Peter Suber. Open Access to Science and Scholarship, InfoPaper, an anthology produced for the December 2003 meeting of the U.N. World Summit on the Information Society. <http://www.earlham.edu/~peters/writing/wsis.htm>
- Raman Nair, R. Developing Digital Libraries: Need For Proper Strategies. *Advances in Library and Information Science*. V5. Ed by D C Ojha and D V Kothari. Jodhpur, Scientific Publishers, 2005. Ch 21, P 259-271.
- Ravindran Asari, K; Hussain, K.H and Raman Nair, R. 2002. Nitya Archives: Innovative blending of techniques for selective access to information from digitally organized text (SAIDOT). In: Parthan, S (Ed.) *Proceedings of the National Conference on Information Management in e-Libraries*, 26-27 February 2002 IIT, Kharagpur. Allied Publishers Limited, New Delhi, 275-285.
- Sathikumar, C. S. and Raman Nair, R. and Bhagi, N. K. Digital Archive of Kerala Legislative Assembly Proceedings. 2007 [Report]. <http://eprints.rclis.org/9296/>
- Sathikumar, CS and Rajan, PD. Digital Archive of Kerala Legislative Assembly Proceedings: A Unique Model in Indian Context. Trivandrum, KLAL, 2009.
- Suber, Peter and Raman Nair, R. and Hussain, K. H. Open Access to public funded research: a discussion in the context of Mahatma Gandhi University digital archives of doctoral dissertations. 2009. In CALIBER 2009, Pondicherry (India), 25-27 February 2009. [Conference Paper]