Information Studies, Vol. 11, N o. 1, July 2002

Creation of Digital Archives in Indian Languages Using CDS\ISIS: Development of M-ISIS (Malayalam ISIS) and 'Nitya'

K H. Hussain, P. Vijayakumaran Nair, R. Chitrajakumar, K. Ravindran Asari, R. Raman Nair

Need for developing Information Systems in Indian vernacular languages is now felt more than ever. Use of local scripts becomes inevitable in the creation of Information systems for digitized palm leaves, manuscripts and local documents. M-ISIS is a localized version of CDS\ISIS for Malayalam language, which is programmed using ISIS32.DLL created by UNESCO. It is a bibliographic information system for Malayalam documents using Malayalam script. Together with the techniques of 'Nitya Digital Archives' it searches using Malayalam script and retrieves both references and full texts. A special font 'Meera.ttf' is designed and a rendering engine to process conjunct formation is programmed, enabling both data entry and query formulation in Malayalam. Paper describes the creation of a digital archive of 'Mathrubhumi' weekly. A few issues of the weekly were scanned to PDF (Portable Document Format) using Adobe Acrobat. Catalogue data of every individual article was fed into a CDS\ISIS database through a data entry worksheet designed in M-ISIS. The prototype shows immense possibilities of localization of CDS\ISIS in Indian languages

1. Introduction

In 1990s many college and university libraries in Kerala created electronic catalogues of Malayalam books using English scripts. Databases were created using packages like CDS\ISIS and LibSys, but all the attempts failed to yield expected results. Since most of the library 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. Some of the libraries devised transliteration schemes, but Malayalam data became so un-natural that even the library professionals could not decipher them. (e.g. 'Kerala' became 'Kaerhalha'). 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.

2. Information Systems in Local Languages

2.1 GIST

The GIST technology devised by C-DAC was the only solution to circumvent the situation. The technology was based on 'Key board hooking', an interface that enabled the data input in any running application (word processors, DTP packages, DBMS, etc) using any Indian script. LibSys advocated the installation of GIST along with their packages. Data could be entered in Malayalam in their data entry module. But in the search module the interface often failed and the index appeared in unrecognizable Roman characters. After installing Libsys, input of Malayalam data in Calicut University Library was completed in 1998, but the search is an unattainable goal even after six years.

2.2 Embedding Indian Scripts in OS

Real solution lies in embedding Indian scripts in operating systems like MS Windows, Linux, etc. Microsoft has already completed the embedding of Hindi, Tamil, etc, but embedding of Malayalam is delayed for two years after their declaration in 2002. They have released a beta version in June 2004, but far from perfection. Since MS Windows is not an open system, algorithm and codes of their rendering engine (shaping engine named Uniscribe) will remain a secret to them and cannot be studied and altered by an external agency for betterment. Apart from 'Access' or 'MS SQL Server' one should be at their mercy to have a correct 'character formation behavior' with other DBMS like Oracle, MySQL, etc.

Rachana has been successful in designing the first Open True Type Font (OTF) for Malayalam. Stendek R&D, Cochin is trying Malayalam embedding in Linux. They have already started programming the rendering engine for Malayalam. When completed Linux in Malayalam will present a more open, strong and net-workable environment compared to Microsoft's. But it will take a few years to popularize Linux in Kerala and to have library management packages.

3. Attempts in Malayalam

3.1 Rachana

At present there exist at least twenty types of character sets and mappings in Malayalam word processing! This situation, unheard in any other Indian languages was originated from the government initiative in early seventies to modify Original/ Traditional characters of Malayalam for typewriter usage. The typewriter character set is since then known as 'Modified' / 'New' Malayalam script having only one tenth of the number of original characters. By the end of nineties word processing, typesetting and even writing by new generation students became chaotic with the mixing up of Original and Modified characters. In 1999 'Rachana', a linguistic forum started its campaign for the use of Original/Traditional script in Malayalam computing with ardent supports from most of the scholars, writers and linguisticians in Kerala. Rachana argued that solution to the present confused state of Malayalam computing lies in the use of Original/traditional characters. A word processor called 'Rachana' was programmed to show the feasibility of Original character set. Since then a lot of books including Malayalam Bible and Ramayana were typeset and printed using Rachana's fonts. During these five years Rachana movement has gained considerable momentum that following the example of Linux, Microsoft is also planning to embed an Open True Type Font (OTF) based on the Original charactersof Malayalam.

Considering the present trend set up by Rachana, we have adopted the Original character set of Malayalam for the localization of CDS\ISIS to create M-ISIS. A unique font called 'Meera.ttf' is designed and a rendering engine is programmed to process conjunct formation in Malayalam by the method set up by Rachana.

3.2 Nitya Digital Archive

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 scanned images. 'Nitya' explores 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 (Malayalam ISIS) is an attempt in this direction.

3.3 Brennen CD

The first product using M-ISIS is 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 details were fed in to a database 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 shows the necessity of building up information systems of Malayalam documents using Malayalam script. The use of Original Malayalam script advocated by Rachana shows that it is the most standardized and comprehensive character set, and hence advisable for creating information systems in Malayalam. Achieving search and retrieval using Malayalam script, next step was to combine the technique of M-ISIS and 'Nitya' to create a digital archiving system.

3.4 Mathrubhoomi Weekly Archives

'Mathrubhoomi' weekly published since 1923 is one of the important literary and cultural magazines in Malayalam. A few issues of the weekly were scanned to PDF (Portable Document Format) using Adobe Acrobat. Catalogue data of every individual article was fed into a CDS\ISIS database through a data entry worksheet designed in M-ISIS. Rendering engine for Meera font (i.e. processing mechanism for conjunct formation in Malayalam) was also programmed with M-ISIS. 'Nitya' performs search and retrieval of the full text.

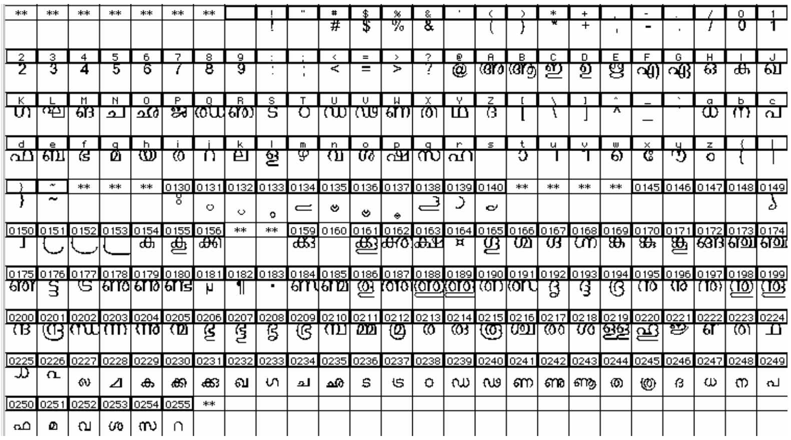
5. M-ISIS and 'Nitya': Implementation Details

5.1 M-ISIS and 'Nitya'

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

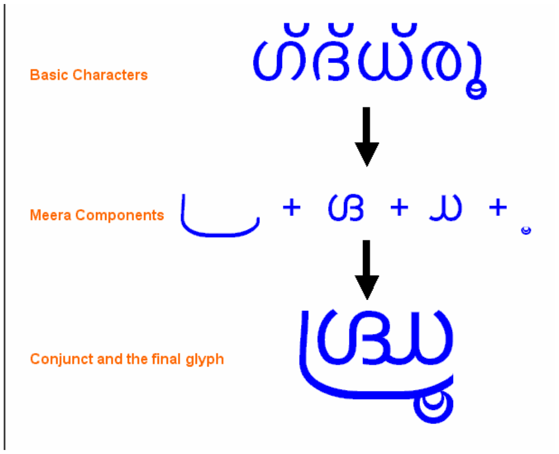
5.2 Font Meera

The font Meera is a True Type Font (TTF) specially designed for M-ISIS using FontoGrapher 4.1. Its character set and glyphs are inspired by Rachana font, the exhaustive Original/Traditional characters of Malayalam.



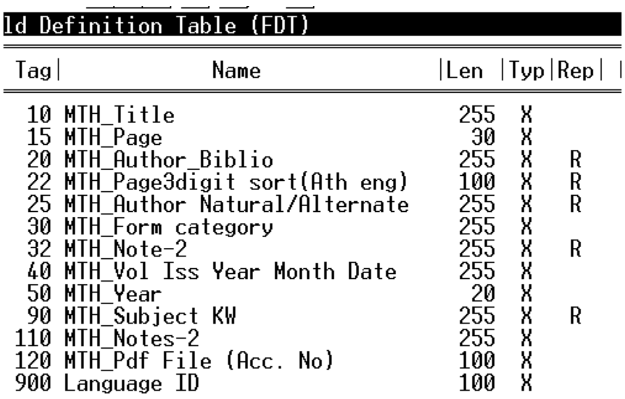
5.3 Rendering Engine

'Inscript' keyboard standardized by DoE for all Indian languages is selected for inputting data. When basic characters are entered M-ISIS combines components in the font Meera to form the final glyph of the conjuncts. This rendering process eliminates the need for 'outer mechanism' like GIST. Since the conjuncts and word formations are inbuilt with M-ISIS the same process can be used for displaying index (dictionary terms) and formulating queries.



5.4 FDT (Field Definition Table) of the Database

For MTH (Mathrubhomi) database 13 fields are defined using CDS\ISIS. Tag 120 is for the file name of scanned text. In our case it is the PDF file name.

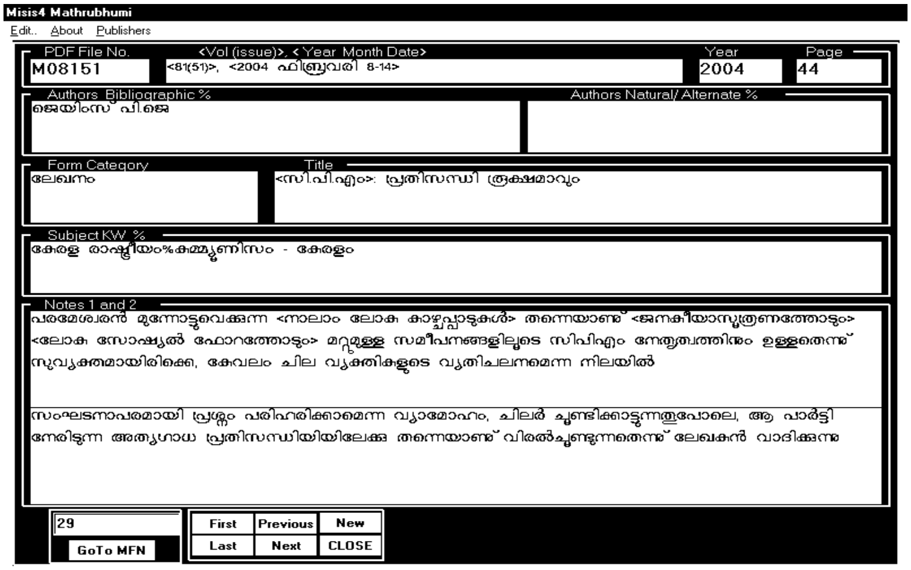


ld Definition Table (FDT)

Tag	Name	Len	Typ	Rep
10	MTH_Title	255	X	
15	MTH_Page	30	X	
20	MTH_Author_Biblio	255	X	R
22	MTH_Page3digit sort(Ath eng)	100	X	R
25	MTH_Author Natural/Alternate	255	X	R
30	MTH_Form category	255	X	
32	MTH_Note-2	255	X	R
40	MTH_Vol Iss Year Month Date	255	X	
50	MTH_Year	20	X	
90	MTH_Subject KW	255	X	R
110	MTH_Notes-2	255	X	
120	MTH_Pdf File (Acc. No)	100	X	
900	Language ID	100	X	

5.5 Worksheet for data entry

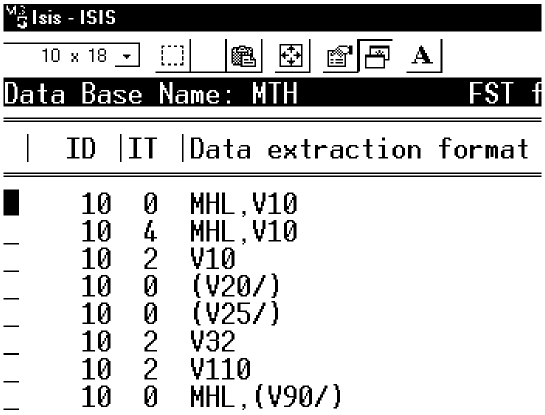
New records are created by clicking 'NEW' button. Provisions are put for modifying records already entered (FIRST, LAST, PREVIOUS, NEXT buttons).



PDF File No.: M08151
<Vol (issue)>, < Year Month Date>: <81(51)>, <2004 ഫിബ്രവരി 8-14>
Year: 2004
Page: 44
Authors Bibliographic %: ജെയിംസ് പി.ജെ
Authors Natural/ Alternate %:
Form Category: ലേഖനം
Title: <സി.പി.എം>: പ്രതിസന്ധി രൂക്ഷമാവും
Subject KW %: കേരള രാഷ്ട്രീയം%കമ്മ്യൂണിസം - കേരളം
Notes 1 and 2:
പരമേശ്വരൻ മുന്നോട്ടുവെക്കുന്ന <നാലാം ലോക കാഴ്ച്ചപ്പാടുകൾ> തന്നെയാണ് <ജനകീയാസൂത്രണത്തോടും> <ലോക സോഷ്യൽ ഫോറത്തോടും> മറ്റുമുള്ള സമീപനങ്ങളിലൂടെ സിപിഎം നേതൃത്വത്തിനും ഉള്ളതെന്ന് സുവ്യക്തമായിരിക്കെ, കേവലം ചില വ്യക്തികളുടെ വ്യതിചലനമെന്ന നിലയിൽ
സംഘടനാപരമായി പ്രശ്നം പരിഹരിക്കാമെന്ന വ്യാമോഹം, ചിലർ ചൂണ്ടിക്കാട്ടുന്നതുപോലെ, ആ പാർട്ടി നേരിടുന്ന അത്യഗാധ പ്രതിസന്ധിയിയിലേക്ക തന്നെയാണ് വിരൽചൂണ്ടുന്നതെന്ന് ലേഖകൻ വാദിക്കുന്നു
29

5.6 FST (Field Select Table) for Indexing

Every word in the title, every phrase in the title bracketed in <....>, every author, every keyword, every phrase in note section bracketed in <...> are indexed.



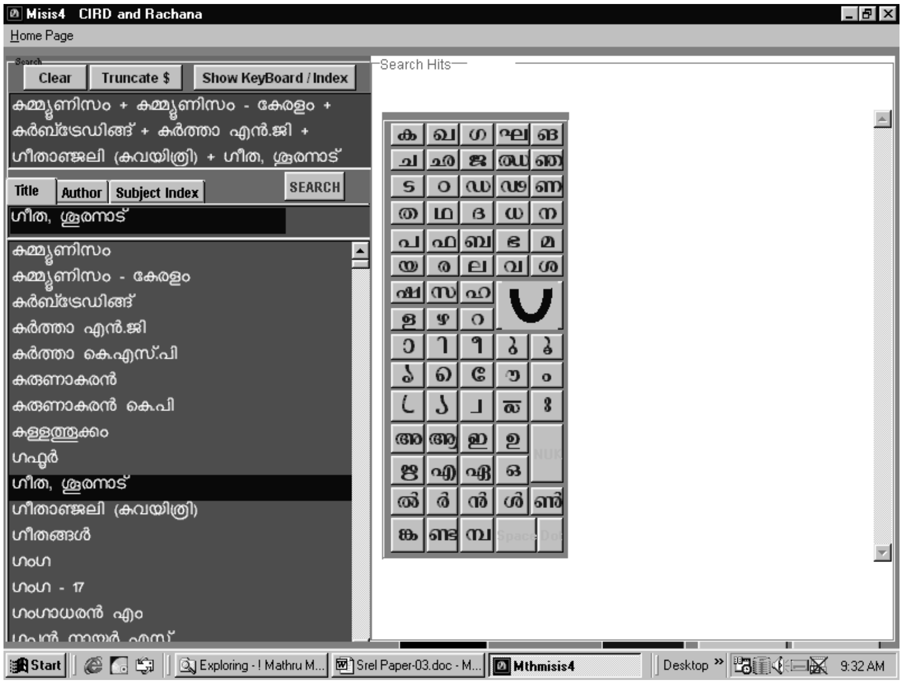
```
Data Base Name: MTH          FST f

 |  ID |IT |Data extraction format

█    10  0  MHL,V10
_    10  4  MHL,V10
_    10  2  V10
_    10  0  (V20/)
_    10  0  (V25/)
_    10  2  V32
_    10  2  V110
_    10  0  MHL,(V90/)
```

5.7 Search and Retrieval

A Malayalam keyboard is simulated, using which a few characters are clicked which shows the terms after these characters in the index box (Dictionary box). Terms are clicked to transfer the term to query area. Queries can later be refined by Boolean operators (+ for OR, * for AND, ^ for NOT).



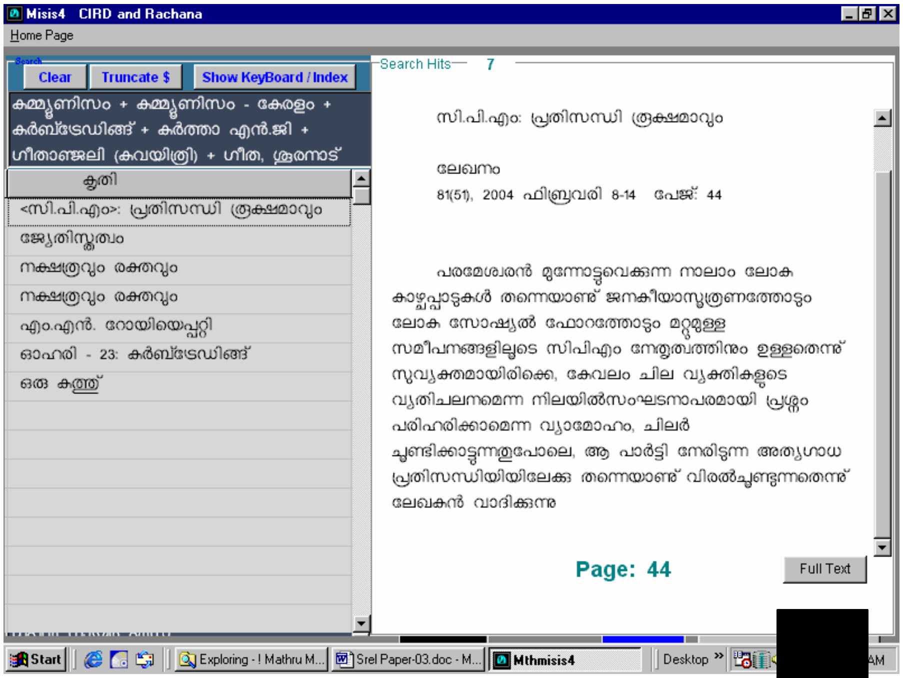
കമ്മ്യൂണിസം + കമ്മ്യൂണിസം - കേരളം + കർബ്ട്രേഡിങ്ങ് + കർത്താ എൻ.ജി + ഗീതാഞ്ജലി (കവയിത്രി) + ഗീത, ശൂരനാട്

ഗീത, ശൂരനാട്

കമ്മ്യൂണിസം
കമ്മ്യൂണിസം - കേരളം
കർബ്ട്രേഡിങ്ങ്
കർത്താ എൻ.ജി
കർത്താ കെ.എസ്.പി
കരുണാകരൻ
കരുണാകരൻ കെ.പി
കള്ളത്തുക്കം
ഗഫൂർ
ഗീത, ശൂരനാട്
ഗീതാഞ്ജലി (കവയിത്രി)
ഗീതങ്ങൾ
ഗംഗ
ഗംഗ - 17
ഗംഗാധരൻ എം

Index and Keyboard

Search is performed by clicking the 'Search' button. The retrieved records are first displayed with titles in a table (string grid). Clicking each title, its full reference is displayed and then clicking 'Full Text' button the text is opened in Acrobat reader.



കമ്മ്യൂണിസം + കമ്മ്യൂണിസം - കേരളം + കർബ്ട്രേഡിങ്ങ് + കർത്താ എൻ.ജി + ഗീതാഞ്ജലി (കവയിത്രി) + ഗീത, ശൂരനാട്

കൃതി
<സി.പി.എം>: പ്രതിസന്ധി രൂക്ഷമാവും
ജ്യോതിസ്സത്വം
നക്ഷത്രവും രക്തവും
നക്ഷത്രവും രക്തവും
എം.എൻ. റോയിയെപ്പറ്റി
ഓഹരി - 23: കർബ്ട്രേഡിങ്ങ്
ഒരു കത്ത്

സി.പി.എം: പ്രതിസന്ധി രൂക്ഷമാവും

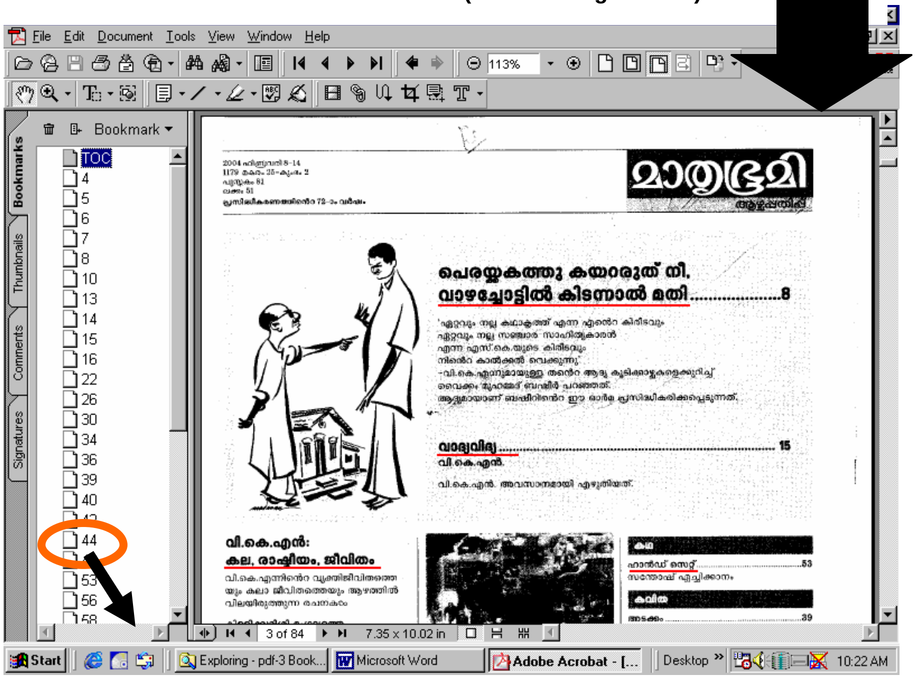
ലേഖനം

81(51), 2004 ഫിബ്രുവരി 8-14 പേജ്: 44

പരമേശ്വരൻ മുന്നോട്ടുവെക്കുന്ന നാലാം ലോക കാഴ്ചപ്പാടുകൾ തന്നെയാണ് ജനകീയാസൂത്രണത്തോടും ലോക സോഷ്യൽ ഫോറത്തോടും മറ്റുമുള്ള സമീപനങ്ങളിലൂടെ സിപിഎം നേതൃത്വത്തിനും ഉള്ളതെന്ന് സുവ്യക്തമായിരിക്കെ, കേവലം ചില വ്യക്തികളുടെ വ്യതിചലനമെന്ന നിലയിൽസംഘടനാപരമായി പ്രശ്നം പരിഹരിക്കാമെന്ന വ്യാമോഹം, ചിലർ ചൂണ്ടിക്കാട്ടുന്നതുപോലെ, ആ പാർട്ടി നേരിടുന്ന അത്യഗാധ പ്രതിസന്ധിയിയിലേക്കു തന്നെയാണ് വിരൽചൂണ്ടുന്നതെന്ന് ലേഖകൻ വാദിക്കുന്നു

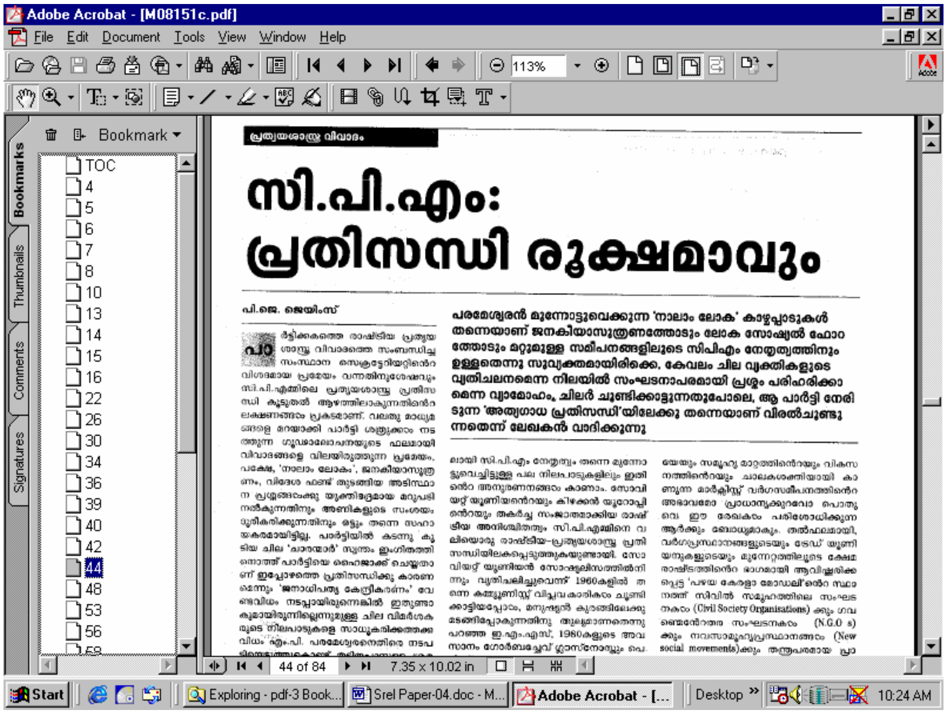
Page: 44

Short and Full Reference (Note the Page No. 44)



Full Text opened in Acrobat Reader. Bookmarks in left

Each issue of Mathrubhoomi weekly is scanned and kept as a single PDF file and is book-marked with page numbers. This page number is displayed more vividly while the full reference of the retrieved article is displayed in M-ISIS. Acrobat Reader opens the document with bookmarks of page numbers. Clicking the page number the respective page is opened. Navigational links are also provided with each item in the table of contents.



പ്രത്യയശാസ്ത്ര വിവാദം

സി.പി.എം: പ്രതിസന്ധി രൂക്ഷമാവും

പി.ജെ. ജെയിംസ്

പരമേശ്വരൻ മുന്നോട്ടുവെക്കുന്ന 'നാലാം ലോക' കാഴ്ചപ്പാടുകൾ തന്നെയാണ് ജനകീയാസൂത്രണത്തോടും ലോക സോഷ്യൽ ഫോറത്തോടും മറ്റുമുള്ള സമീപനങ്ങളിലൂടെ സിപിഎം നേതൃത്വത്തിനും ഉള്ളതെന്നു സുവ്യക്തമായിരിക്കെ, കേവലം ചില വ്യക്തികളുടെ വ്യതിചലനമെന്ന നിലയിൽ സംഘടനാപരമായി പ്രശ്നം പരിഹരിക്കാമെന്ന വ്യാമോഹം, ചിലർ ചൂണ്ടിക്കാട്ടുന്നതുപോലെ, ആ പാർട്ടി നേരിടുന്ന 'അത്യഗാധ പ്രതിസന്ധി'യിലേക്കു തന്നെയാണ് വിരൽചൂണ്ടുന്നതെന്ന് ലേഖകൻ വാദിക്കുന്നു

Page 44 accessed through Bookmarks

6. Conclusion

For creating a local version of CDS\ISIS in Indian languages one should develop:

? A new font or better adopt an existing font

? A front end using ISIS32.DLL that should have the following two modules

? Data entry module with a data entry worksheet that can handle conjunct formations in the local language

? Search module that can formulate queries in the local language

This is the first time a bibliographic full text retrieval system using Malayalam script is developed. M-ISIS shows immense possibilities of localization of CDS\ISIS in Indian languages. By giving provision to open an external PDF file the reference system is transformed in to a full text retrieval system. CDS\ISIS can be effectively utilized to develop and distribute digital archives of local documents in Indian languages. Both CDS\ISIS and Acrobat reader are free and hence distribution doesn't pose any legal problems. Faster retrievability and easier portability of CDS\ISIS make it ideal for CD publishing of full texts.

References

Hussain, K H; Raman Nair, R and Raveendran Asari, K 2002. Importance of search and retrieval in CD-ROM full text publishing: Experiments using PDF documents and ‘Nitya’ archival system. *Information Studies* 8(3): 173-180

Ravindran Asari, K. 2000**.** Nitya Archives: A Solution for Organizing Digitized Content in Agriculture**.** Agricultural Information Systems: Vision 2020. Thrissur, IASLIC Study Circle, 2000.

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.

Sulochana Devi, L. Information System for Research in Sanskrit and Indology. In the Proceedings of the National conference on Vedic Sciences. Madras, JGRCVS, 1997