Towards a Total Internet Solution for the Tamil Language

Through Singapore Research

Naa Govindasamy

Lecturer

National Institute of Education (NIE)

Nanyang Technological University (NTU)

469, Bukit Timah Road

Singapore 259756

http://www.irdu.nus.edu.sg/tamilweb

email: govind@cir.nus.edu.sg

(This Paper was presented at the SAARC Conference on Extending the Use of Multilingual & Multimedia Information Technology at Pune, India, on September 1 –4 1998. The Conference was organised by the Centre for Development of Advanced Computing (C-DAC), a Scientific Society under the Department of Electronics (DoE) New Delhi, Government of India.)

(The author wish to thank Dr Tan Tin Wee, Associate Director for the Centre for Internet Research (CIR), National University of Singapore and Mr Leong Kok Yong, Research Officer of CIR, for giving valuable advice in the preparation of this paper.)

Abstract

Internet revolution has enabled the widespread dissemination of information throughout the world. Most of the content is in Romanized characters. Research is going on in some countries to enable non-Roman scripts to accessible on the Internet. This paper will discuss and demonstrate how, through a successful research collaboration in Singapore, Tamil language content is now freely accessible, searchable, conveniently emailable and easily composed and edited on the Internet through all three popular platforms Unix, PC and Mac.

1. The Need for A Multilingual Internet in Singapore

Singapore is a multilingual and a multiracial country. English, Chinese, Malay and Tamil are the official languages. Most of the government and public documents are published in these four official languages. However, until recently, it was not possible for the Chinese and Tamil languages to be disseminated through the World Wide Web on the Internet. In 1994 Dr Tan Tin Wee, my research collaborator initiated work in this area while he was the head of Technet Unit, the first Internet service provider for Research and Educational Institutions in Singapore. Technet Unit was directly under the supervision of the Computer Centre of the National University of Singapore. (Technet Unit has since been commercialised to become Pacific Internet, one of the three ISPs for Singapore. The others are Singnet and Cyberway.)

By mid-1994, Technet Unit initiated the Singapore INFOMAP project which provided a one-stop WWW home page for Singapore. He wanted the four official languages to be represented in the INFOMAP. Since English and Malay are using the Roman script, diplaying these two languages on the WEB was not a problem. By the end of 1994, Technet had successfully implemented an Experimental Chinese WEB server in Singapore. So the problem of displaying the Chinese script on the Web was solved.

However, displaying Tamil script on the Web, and communicating through Tamil on Internet was a problem.

There was no Tamil Information System on the Internet which provides a display system in Tamil and English simultaneously on the Text Mode using a Tamil-English single font file. There were a few servers, which were providing Tamil script using GIF image files. Tamil Eelam Page (http://www.eelam.com) was and is still very active in this direction. Tamil Nadu Home Page, and Tamil Electronic Library (http://www.geocities.com/Athens/5180/index.html) are other popular Tamil Web Sites on Internet at that time. Tamil Electronic Library was using (and is still using) a mono 7bit font (Mylai) for the Tamil display on the Web.However, Mylai font cannot support native emailing at that time. So there was a need to develop a Tamil Internet System which should go beyond Web display.

In May 1995, I met Dr Tan at the Technet Unit, National University of Singapore, soon to become the Internet Research and Development Unit (IRDU) (now upgraded to Centre for Internet Research - http://www.cir.nus.edu.sg). We identified the potential solutions and agreed for a possible research collaboration between NUS and my institution, NIE, NTU, the two instititutions of higher learning in Singapore at that time. (http://www.irdu.nus.edu.sg) & (http://www.nie.ac.sg)

At that time, on my own, I was in the process of developing True Type fonts and a Tamil software for Windows Applications (Kanian™ Tamil Software).

As the service provider arm of the Technet was sold to a private consortium and renamed Pacific Internet in September 1995, the nascentTamil Internet Research was inherited by the newly formed Internet Research and Development Unit (IRDU). This Research Unit was funded by the National Science and Technology Board (NSTB) (http://www.nstb.gov.sg) of Singapore. Mr Leong Kok Yong, just graduated from the Nanyang Technological University, joined IRDU, and became one of the key member for the TamilWEB project.

2. Objective

During the Technet period, when Dr Tan and I, conceptulised the TamilWEB (http://www.irdu.nus.edu.sg/tamilweb) Project. We had a very clear objective. That is:

1. ***to develop a bilingual font system for The Total Internet Communication in the Tamil language.***

That means, the system that we intended to develop:

a) should provide display of Tamil & English Text simultaneously on the Internet Applications. (Web Browsers & Email sofware packages.)

b) Tamil and English text should be easily communicated and retrieved in Plain Text.

c) should work across Platforms. (PC, Mac & Unix)

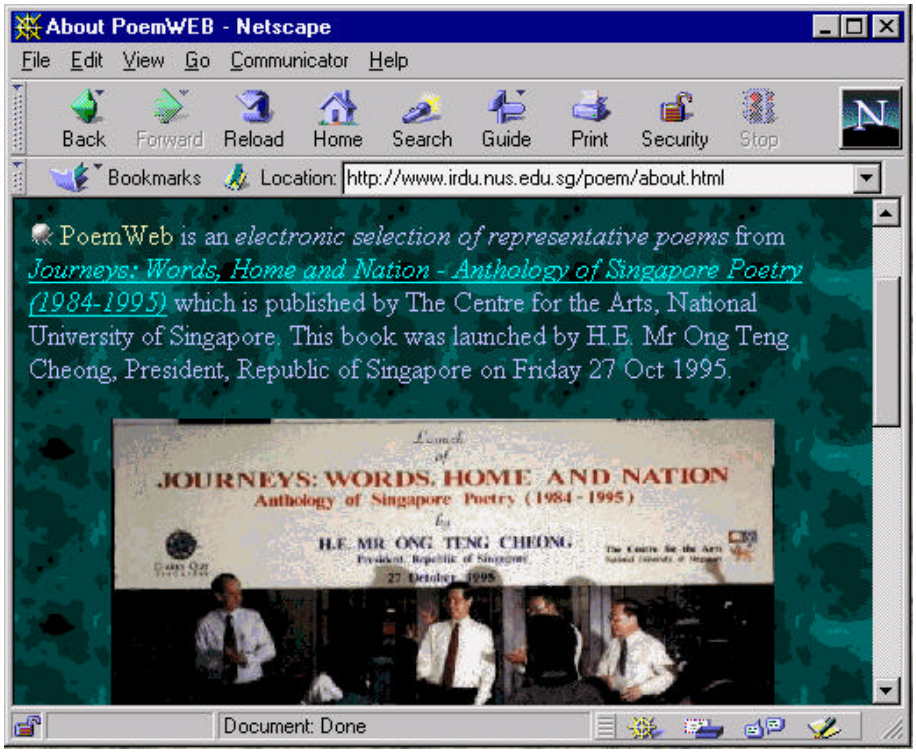
d) should be searchable in Tamil

e) should let the user type Tamil in the Web browser's Forms, and the typed word should be seen in Tamil.

f) Should allow the user to read Tamil & English in terminal emulation mode (telnet).

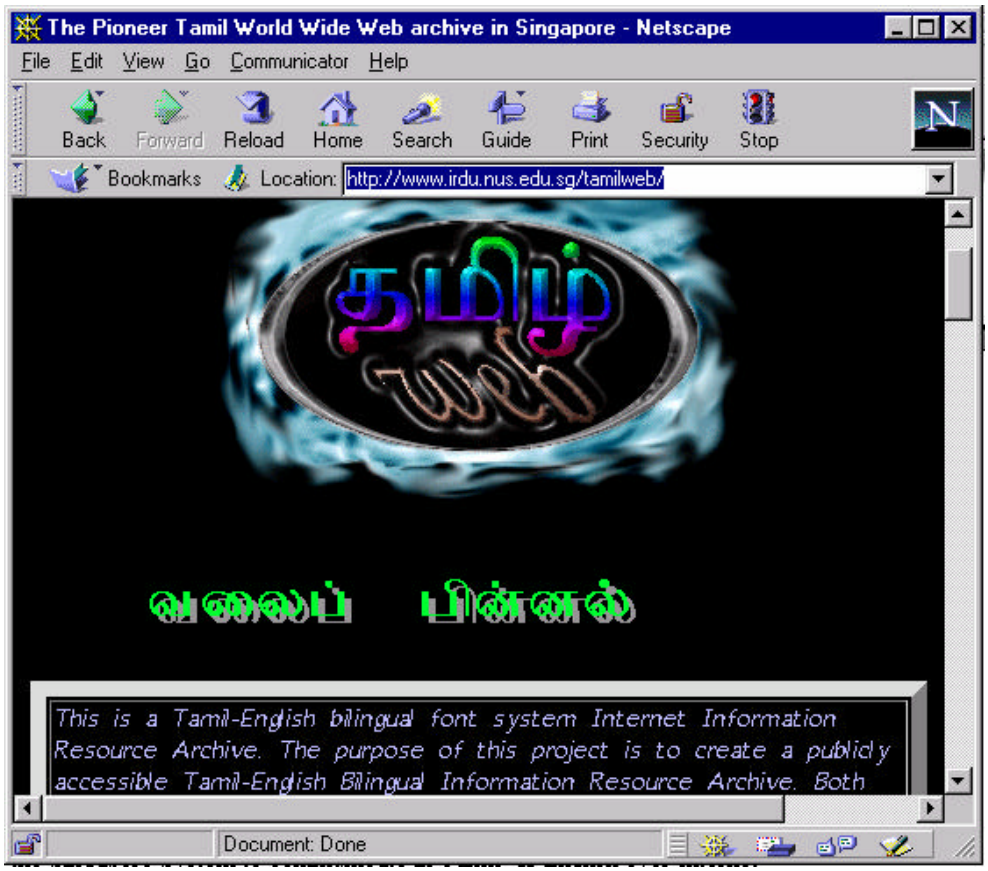
3. Prototype Testing, Preview and Official launch

The prototype of our system was tested during the launch of PoemWEB. PoemWeb (http://www.irdu.nus.edu.sg/poem) is an electronic selection of representative poems from the four official languages, from the book , Journeys: Words, Home and Nation - Anthology of Singapore Poetry (1984-1995) which was published by The Centre for the Arts, National University of Singapore. This book was launched by H.E. Mr Ong Teng Cheong, President, Republic of Singapore on Friday 27 Oct 1995.



The preview of the first phase of TamilWEB project was shown to the public and to the Press at the "Internet for Everyone – 1995" at Suntec City Exibition Hall during 1 –3 December 1995.

TamilWeb was officially launched by the Honourable Member of Parliament, Dr Ong Chit Chung (Chairman, GPC for Education and MP for Bukit Batok) on 2 February 1996. (http://www.irdu.nus.edu.sg/tamilweb)



Since then the Tamil language teachers in Singapore and the Internet Users from locally and abroad, are using the Singapore System to communicate in Tamil & English over Internet and have created a significant volume of bilingual Web pages in Tamil and English.

niversity of Pennsylvania's Penn

ilingual Website for Learning and
iage Teaching and Learning, with the
nia.

The ruling party of Tamil Nadu Dravida Munneetra Kazhagam (DMK) Website is another important site, using our system. (http://www.thedmk.org)

4. Purpose of this paper

This paper will try to explain and demonstrate, how Tamil, one of the Indian languages, has achieved the Total Internet Solution, through Singapore Research. Most of my presentation will be done through Internet.

ohical Web browsers on the Internet, was the most preferred, Text-mode vailable at that time. Content on the Content in non-Roman languages was

5. Tamil Internet tools from Singapore

These Internet Tools originating from our research and software development are free for downloading:

1. TamilNet.ttf (PC propotional font)

2. TamilFix.ttf (PC fix width font)

3. Tamilnet.hqx (Mac propotional font)

4. Tamilfix.hqx (Mac fix width font)

5. Tamilnet18.bdf (Unix font)

6. Tamilfix.bdf (Unix Font)

7. Tamil Keyboard Manager (for PC)

8. Tamil Keyboard Manager (for Mac)

9. Xkeymap (Tamil input system for UNIX)

10. Mirage (CGI Application Software for rendering Multilingual encoding text into GIF images for display on web browser.

11. Applet input sysytem for Tamil word search

For more information, please consult the URL - http://www.cir.nus.edu.sg/tamilweb

6. Font encoding

The key tool for the project was the creation and design of a Bilingual Tamil-English single font system.

We have designed a bilingual font set for the display of both Tamil and English simultaneously. This was done by making use of the upper extended ASCII character range for the Tamil characters, while retaining the basic English alphabet and punctuation intact in the lower ASCII range. This will allow most of the Web world in English (or other Romanized languages) to be traversed; at the same time, Tamil codes will be recognized and displayed correctly when they occur, without having to change font set. Figure below shows the character map for the Tamil-English font set.

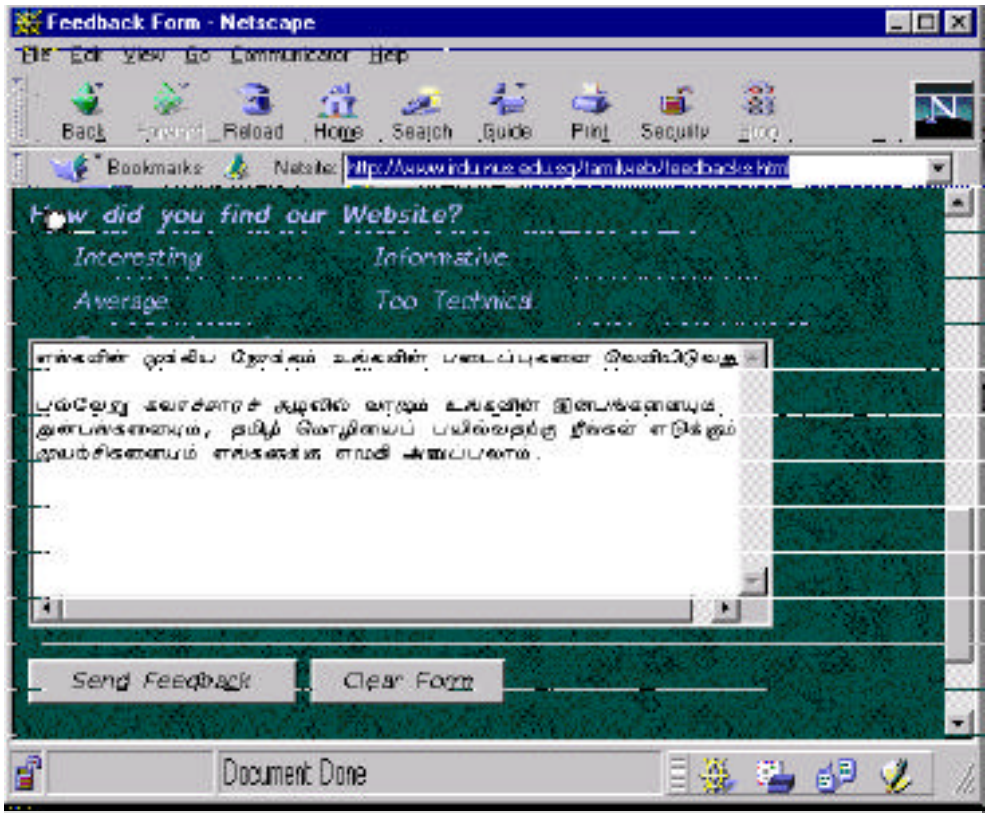
One important point to note is that the upper ASCII portion does not have enough code space to include all the possible Tamil character glyphs (>200). As such, we made use of the kerning feature built into the Postscript and the True-Type font technology to combine two Tamil characters into a new character glyph not found in the above ASCII table. With the combination of two simpler character glyphs to give a more complex glyph, we can then include the entire Tamil character set within one single font, together with the English alphabet. To allow users to input these Tamil characters, a corresponding keyboard layout mapping has been devised by mapping the keys on a normal English (QWERTY) keyboard to the extended ASCII range where these Tamil characters reside. A toggle key enables the user to switch between the two modes.

"Tamilnet " propotional font was developed to display Tamil & English on the Web browser. However, the variable proportional font cannot be viewed in the Web browser's Forms.

For this purpose, a fixed width font, "Tamilfix" was developed. This font is very similar to the "Courier" font. The "Tamilfix" font is simply doing the work of the "Courier" font. Only with this fixed width font, Tamil can be typed in the Web browser Forms. In the Web browser, the form filling feature is a very important component for interactivity. If the user wish to communicate in Tamil to webmaster or the author of the webpage, he or she has to type Tamil into the Form.

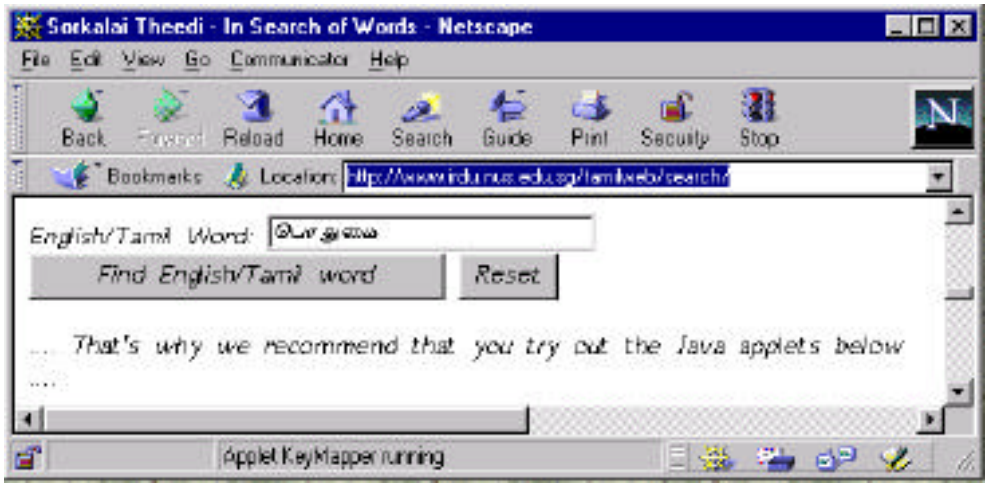
7. Keying in the Tamil characters in the Forms

When the user is keying in Tamil script in the Form, the Tamil Characters should appear on the Forms as they are typed for immediate visual feedback. Only then can meaningful communication and interactivity take place. We achieved this through the "Tamilfix" font and the keyboard input system.

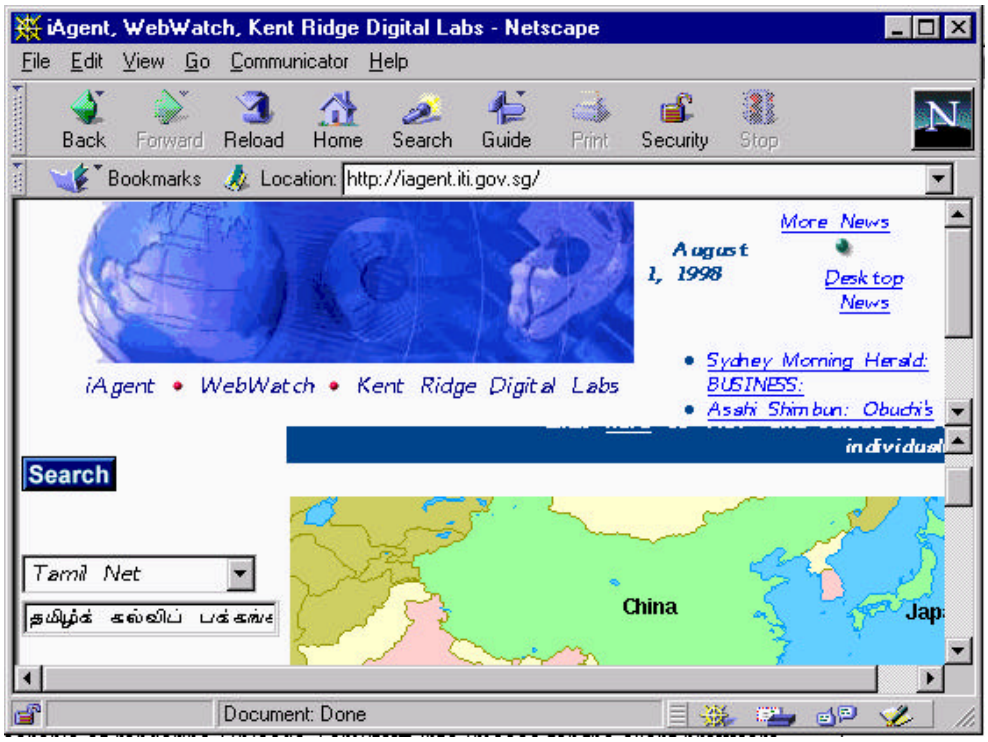


Another important factor in any database creation is the Search function. If a Search function is not possible in a particular system, creating a database, is out of the question. When we launched the TamilWEB on 2.2.1996, we demonstrated the Search function. In the search form, Tamil words were keyed in for searching against a database of Tamil text.

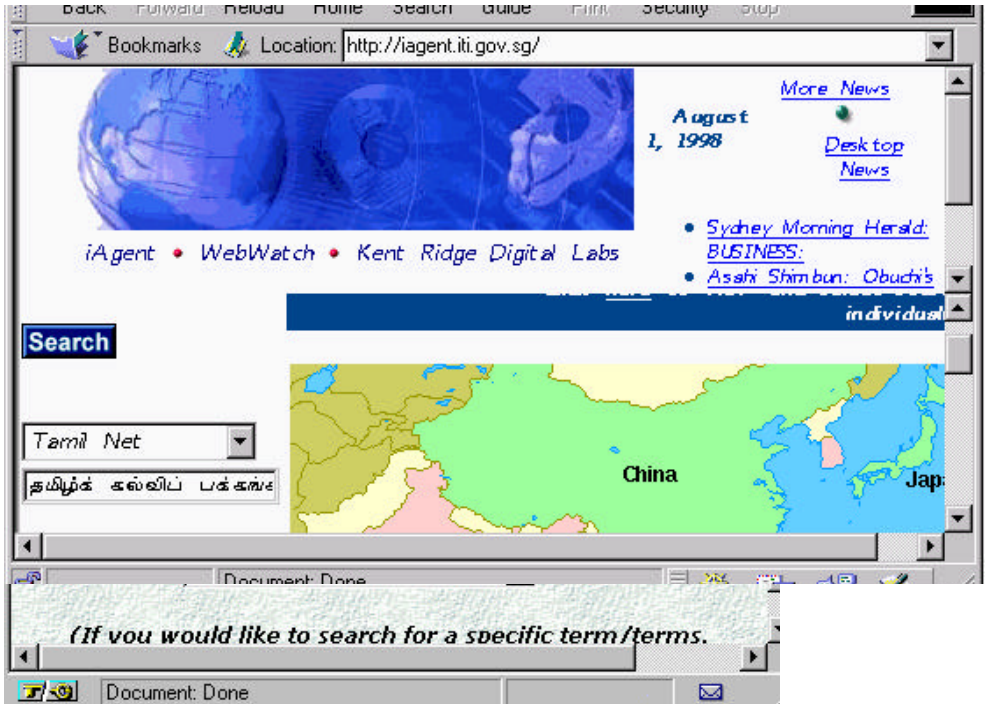
For search and retrieval, the submitted string in extended ASCII for Tamil (and in English as well for bilingual searches) is parsed by the httpd server and submitted as a search string to any indexing engine that has multilingual capability. In the case of Tamil, we used a simple WAIS-SF indexer and demonstrated the utility. Hits were returned in the same encoding, and displayed in the same way as described above, with bilingual capability. In fact, this powerful search function is taking place across the various platforms.



In the Singapore Government Web site (htt://www.gov.sg), searching for Tamil keywords by typing Tamil script is possible. The IAgent (http://iagent.iti.gov.sg)search engine will deliver the results in the form of webpages, using Singapore Tamil font encoding.

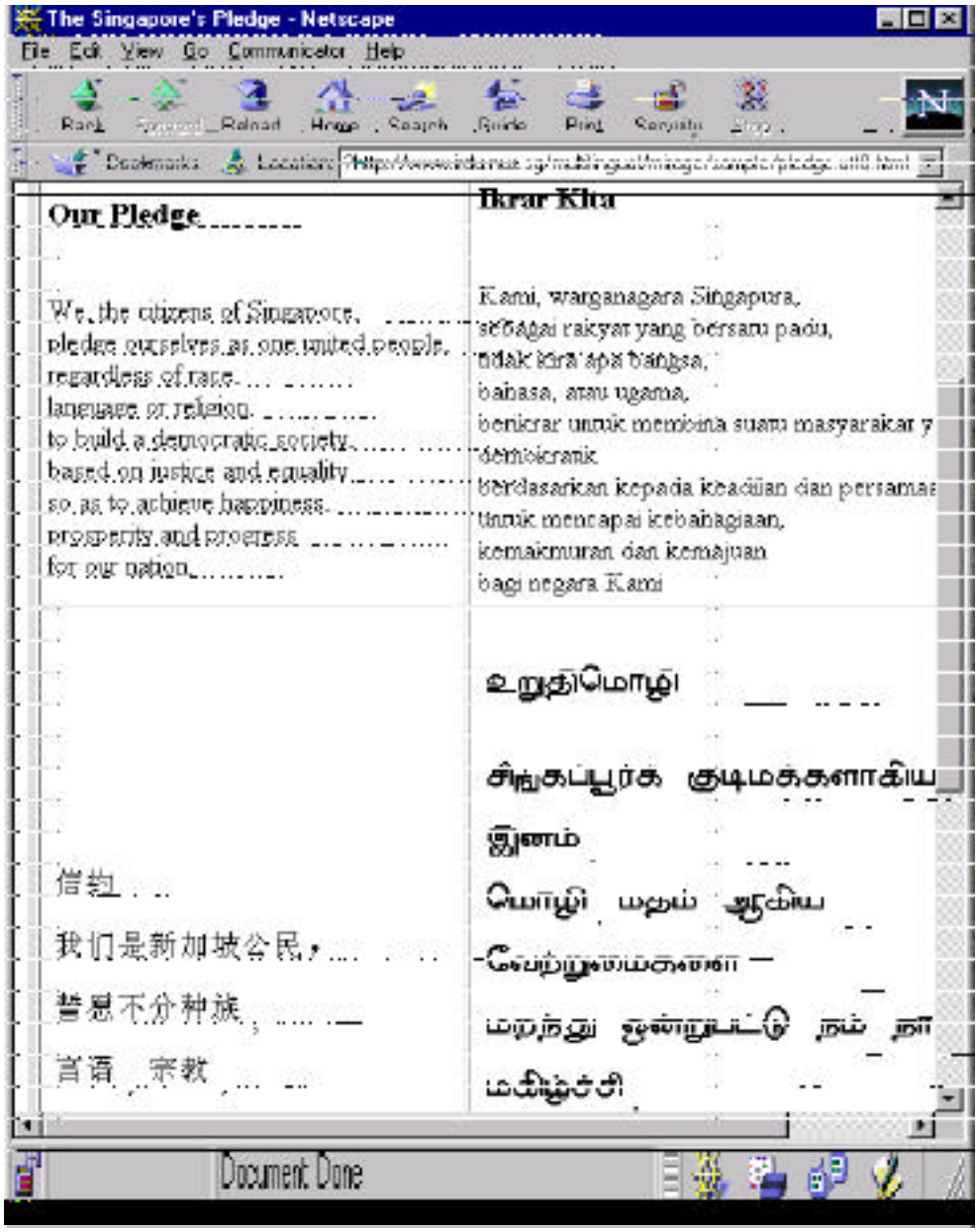


Ministry of Education's Search Engine in Tamil



However, in some cases, users are unable to use our fonts and encoding system for unknown reasons. In this situation, we have invented another solution.

Our research team has produced a CGI Application software for rendering multilingual encoding text directly into images for display on any web browser as embedded images. It is called Mirage. (http://www.cir.nus.edu.sg/multilingual/mirage) When this application is added on to the server, the server is capable of rendering Unicode Tamiltext into images for the client browsers, without any helper application or any font installation. The significance of this system is that, in the client browser, the user should be able to view multilingual information, originally coded using Unicode.

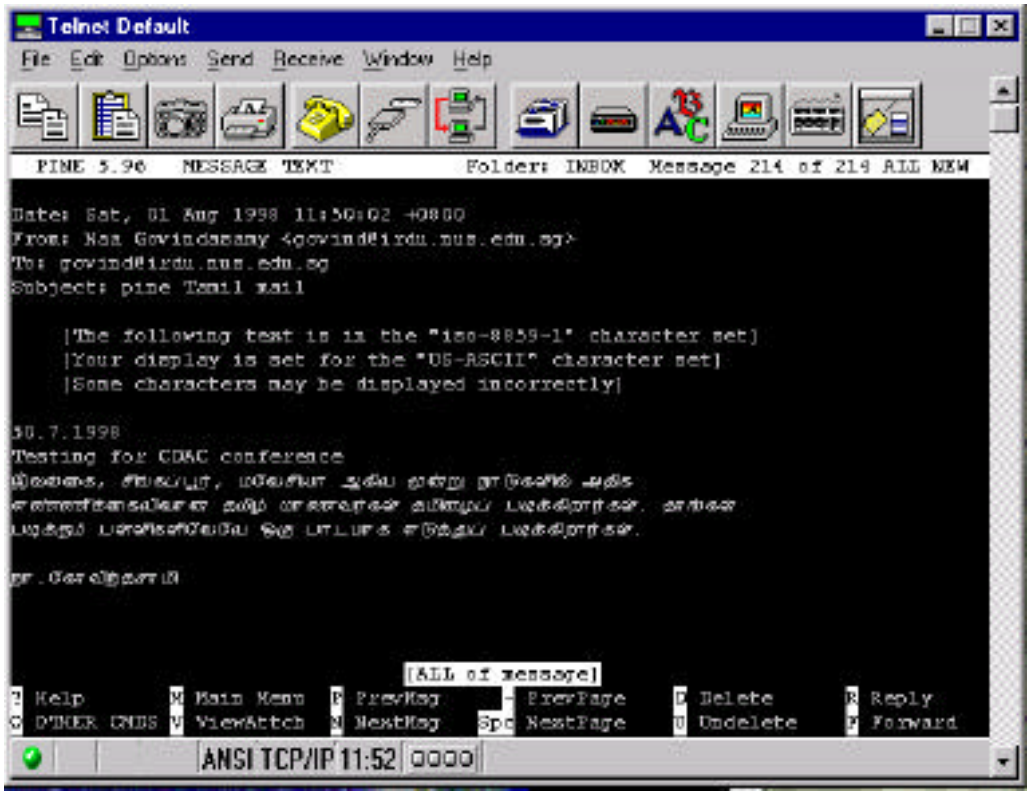


For that matter, any encoding can also be transformed into images using the MIRAGE system, eg. Unicode, ISCII, Kanian-Tamilnet etc. simply by modifying the code table mappings to character glyphs.

Now, I will be demonstrating another of the important feature of our system.

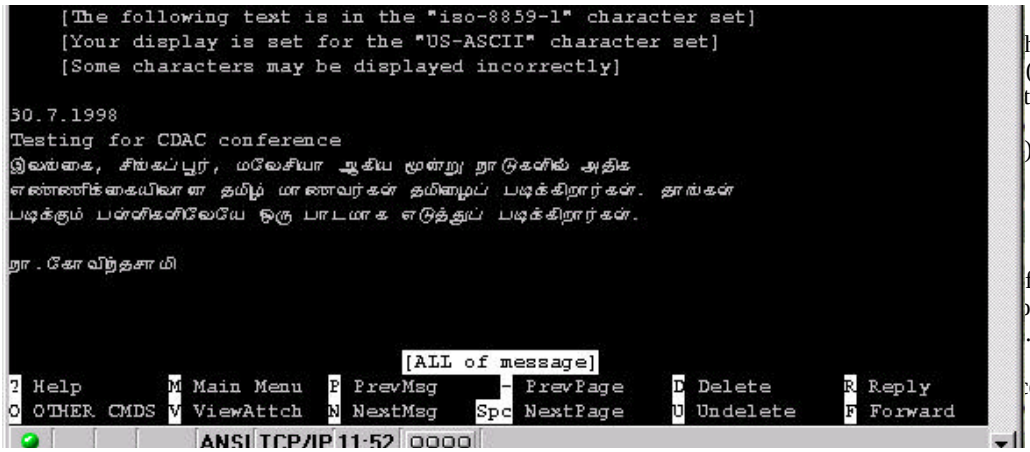
8. Viewing Tamil on PC Terminal Emulation

When we developed the "Tamilfix" font, we knew that it will make Tamil readable in the PC Terminal emulation (eg Telnet). A Shell access user can read Tamil text in the WWW textual browser "LYNX". He can also read Tamil in the Terminal Email software "PINE".



This is a very important development for the Tamil language. In many developing countries, the number of SHELL access users typically outnumbers the TCP account users. Most users access the internet through a character-based terminal emulation rather than a graphical user interface. As such, our system benefits a lot of SHELL access Internet users. This was made possible with our "Tamilfix" font.

9. Keyboard Input Systems



hambi in India in 1983. This was
(DOE) of India. However, this
the

) (Grimes, 1992), the DOE

frequency of occurrence of the
oard layout was introduced for
. Later

ce at the Anna University, Madras,

The keyboard consists of the 12 basic Tamil vowels placed on the left-hand side of the keyboard and the 18 consonants. The 28 basic vowels and consonants are placed in the lower case of the keyboard, while the 2 least frequently occurring Tamil consonants are placed at the upper case with the 5 Sanskrit sound consonants. For modern Tamil, a vowel will not appear in the

middle or at the end of a Tamil word; it will appear only at the beginning of a word. These basic rules were taken into account when this keyboard layout was designed. The advantage of this keyboard layout is that 99.5 percent of the time, Tamil characters can be typed without pressing the shift key at all. Moreover, the most frequently used vowels and the consonants are placed at the home keys (the middle row of the keyboard). This allows the user to type 68 percent of the Tamil words by using only home keys.

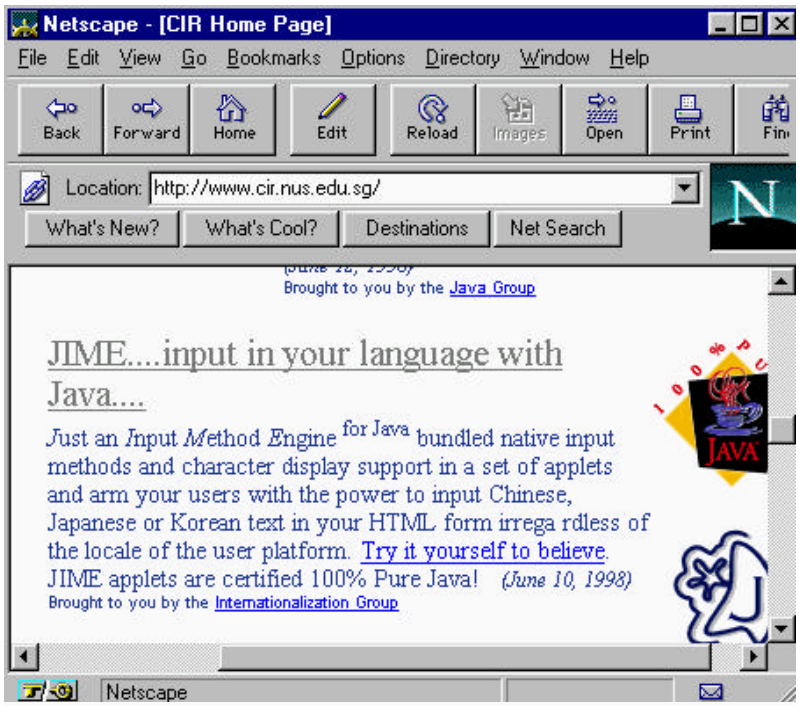
Because of its simplicity and the incorporation of the Tamil grammar, this keyboard layout is very popular in Singapore and Malaysia and has been incorporated in numerous Tamil front-end processing software and word processors, including a commercial version available from the author (Govindasamy, Kanian Bilingual Wordproceesor for PC, Mac & Internet.

<kanian@kanian.com>

The project team has devised a Tamil type writer version for PC & Mac, using KeyMan Keyboard Manager. From our server we are using KeyMan keyboard Manager. For Unix Xkeymap was used to develop the Keyboard Manager for Tamil. These three input devises are downloadable from our Website.

9.1 Toward Java keyboard input systems for total cross platform compatibility using Unicode.

Java Input Method Engine (JIME) for Java from CIR, bundled native input methods and character display support in a set of applets. (http://www.cir.nus.edu.sg/jime) At present the users can input Chinese, Japanese or Korean text in HTML form irregardless of the locale of the user platform.



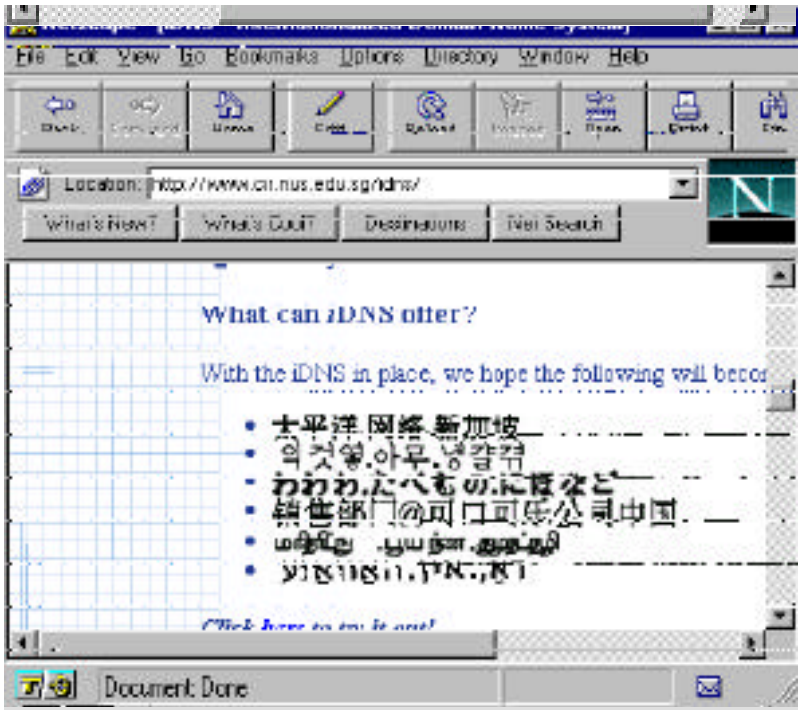
The Tamilweb project team is working towards a Tamil input system in JIME.

10. Future Direction and Conclusion

10.1 Multilingual multiscript URL

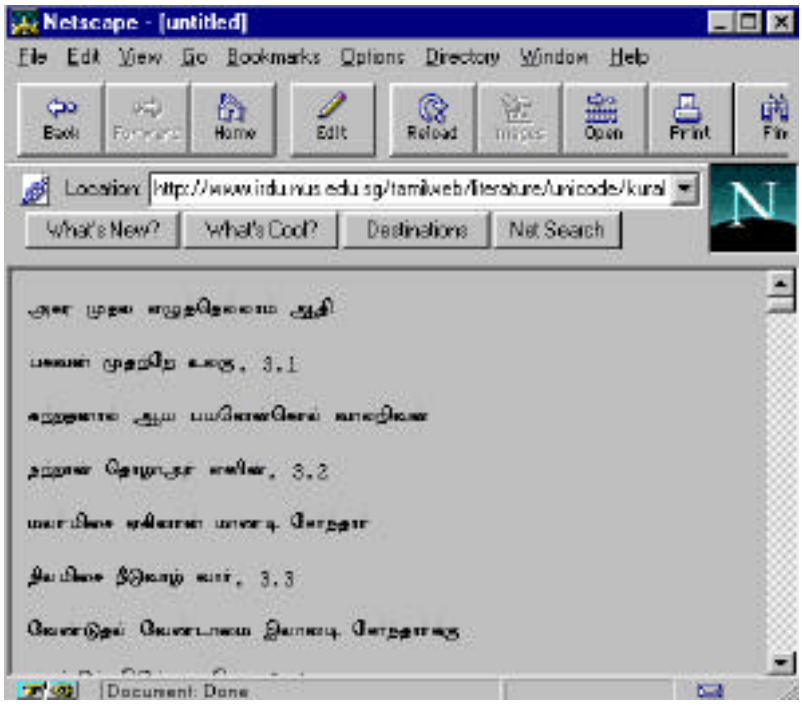
Today, the Internet has reached the four corners of the world to a diverse community with different languages and cultures. The World Wide Web has progressed to address the localization needs of its audience with Web pages in different languages a reality today. However, the Internet Domain Name System (DNS) which started out to be strictly based on a subset of the Latin 1 alphabet, is still mainly English. This restriction also applies to other aspects of the Internet which makes use of domain names as well, e.g. telnet, ftp,email, etc. Now CIR is creating an experimental internationalized DNS as proof of concept that it is viable. It is also creating tools and applications that will enable users to key in URLs in multilingual characters (e.g. Chinese, Japanese, Korean, Tamil etc) It is also designing

a URL forwarding system for multilingual-character URLs.

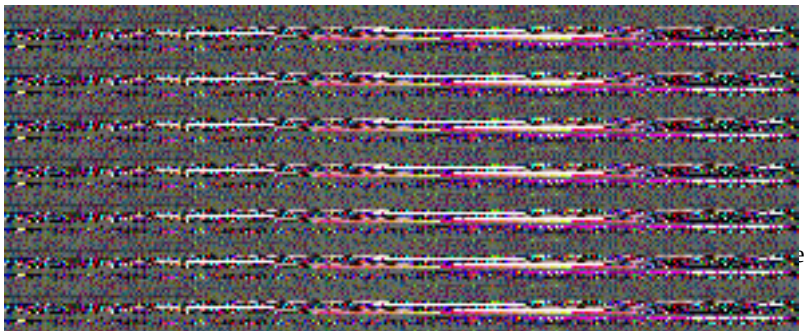


When we achieve multilingual directory and filenames, we will have fully delivered a Tamil script URL in addition to Tamil content on the internet. For the proof of concept visit http://www.idns.apng.org

Now TamilWeb Project is slowly moving to Unicode text archive. The whole Thirukural and part of Purananuru are in two two coding in our server. One in Singapore "kanian/tamilnet" coding. The next coding is the Unicode.



In future most of the digitalized Tamil text in our will be in these two coding. With Multilingual Domain Name System, in future, the Domain name and the URL can be typed in the Tamil language.



e (June, 1996), Multiple Language Support over the

2. Leong Kok Yong , Tan Tin Wee & Lee Teck Chee (March, 1997), Making your Web server render Unicode text for your client users, 10th International Unicode Conference, Mainz, Germany.

3. Bos, Bert (1996). Internationalization/Localization W3C: Non-Western Character Sets, Languages, and Writing Systems. http://www.w3.org/pub/WWW/International/.

4 . Grimes, Barbara F., Editor. (1992). Ethnologue, Languages of the World. 12th Edition. Consulting Editors: Richard S. Pittman and Joseph E. Grimes. Summer Institute of Linguistics, Inc. Dallas, Texas.

5 . Nicol, Gavin T. (1996). The Multilingual World Wide Web.http://www.ebt.com:8080/docs/multilingual-www.html.

6. Govindasamy, N. (1989). New Keyboard for Tamil Computer, by Naa Govindasamy, 7th International Conference of Tamil Studies Seminar Proceedings, Maritius, December 1989.

7. Govindasamy, N. (1994a). Computer and Tamil Teaching. 2nd International Conference of Tamil Language Teaching, Kuala Lumpur, June 1994.

8. Govindasamy, N. (1994b). Kanian Keyboard. Tamil and Computer Conference Proceedings, Anna University, Madras, India, August 1994.