"urumARRi": Online Unicode Conversion of TSCII web page - A Server Side Approach

Sathish Kumar Nithianandam*, Sudhakar Balakrishnan & (Late) Arun Prasath Egambaram
New York, USA

Introduction

This paper addresses needs of Unicode platform users to view TSCII pages without browser setup or font download. Paper provides a solution for TSCII web page publishers to support Unicode users.

"urumARRi" converts TSCII based web page to Unicode page and serves it to the userís browser from Server Side. TSCII web page publishers can add a "urumARRi" location link in their URL for automatic Unicode conversion.

"urumARRi" can also do the reverse, i.e., convert Unicode pages to TSCII web pages for users who are on Unicode non-compatible operating systems like Microsoft Windows 95 and Windows 98.

TSCII web page publishers can extend their support to Unicode users by adding a urumARRi location link with their URL to their pages. Tamil Unicode web page publishers and bloggers can extend their support to TSCII users.

A snapshot of our TSCII to Unicode implementation is shown in Fig. 1: User's browser using "urumARRi" script to view TSCII web page from www.kalkiweekly.com¹ (in Unicode page with Latha Font)

Key Features

* View TSCII web pages without any browser setup or font download in Unicode enabled browsers.
* TSCII web page publishers can extend their support to Unicode users by adding a urumARRi location link with their URL to their pages.
* Enables Search Engines to index TSCII web pages as Unicode pages.
* Multiple online access methods
  * urumARRi Homepage (URL Input box in the page).
  * Using Bookmarklet / Favlet.

¹ www.kalkiweekly.com was chosen to illustrate the parsing of complex HTML by urumARRi.
* "tamizpaTTai" Internet Explorer tool bar.
* Simple implementation, once hosted it is accessible to whole community.
* Server Side solution implemented using Perl v5.8.4 which adds portability and flexibility.
* Highly efficient, low memory footprint code.
* Open source solution for extendibility by Tamil community.
* Tested extensively on Project Madurai TSCII releases & www.infitt.org TSCII web pages.

1. "urumARRi" script on a server serving a TSCII web page, Direct Input URL method is used.; 2. The output web page is identified as Unicode (UTF-8) by Microsoft Internet Explorer browser; 3 & 4. New modified link composed of "ª¬ë±ùÀ" location with the selected link has been shown.

Script Logic
* Retrieves and parses the input HTML from the given URL.
* Converts all TSCII v1.7 characters into Unicode characters (100% accuracy), simple approach. Converts all Unicode characters into TSCII characters for Unicode to TSCII conversion.
* Changes META tag http-equiv argument char-set value from ix-user-defined to utf-8i to hint the browser about the served page type.
* Changes all links in the web page as follows:
  * Converts links to other HTML pages to have the script location with the link.
  * Converts links to frames to support conversion of other frames inside it.
  * Images and links to images are retained as in the original page.
  * Converts relative links to absolute links.
  * Retains English characters as it is.

2 Refer to adjoining paper "tamizpaTTai"
A snapshot of our Unicode to TSCII implementation is shown in Fig. 2: User's browser using urumARRi script to view TSCII web page from www.thoughtsintamil.blogspot.com (in TSCII page with TSC_Avarangal Font)

A snapshot of our Unicode to TSCII implementation for web page publishers using urumARRi script on the same server is shown in Fig. 3: User's browser using urumARRi script to view Unicode web page from www.sathish.net with a link to dynamically convert into TSCII for TSCII users.
Configurable Options

* Convert HTML pages in current server dynamically without changing any relative links (All options including U2T & T2U are currently environment variables).
* Disable changing links to other domains.
* Batch run the script to bulk convert HTML pages in local machine.

Future Scope

This implementation converts TSCII v1.7 to Unicode, can be extended to the other Tamil encodings TAB and TAM.

Limitations

* Web pages from sites with security feature like cookies and https cannot be converted. Due to privacy concerns, form input/POST requests and server side session management are not implemented.
* Rendering of Unicode pages is slower compared to TSCII pages, as Unicode uses more space and memory.

References

TSCII & Unicode standards
http://www.tamil.net/tscii/
http://www.unicode.org/

Conversion Charts
http://www.tamil.net/tscii/faq5.html
http://www.geocities.com/Athens/5180/tscii4.html
http://www.infitt.org/minmanjari/issue2_1/mm-muthu.html

Tested on TSCII web pages
http://www.projectmadurai.org (http://www.tamil.net/projectmadurai/)
http://www.infitt.org/