Unicode-Based Digital Databases for Tamil Cultural Heritage Materials

K. Kalyanasundaram
(Lausanne, Switzerland) and
Muthu Nedumaran
(Kuala Lumpur, Malaysia)

Abstract

With the wide usage of Tamil in native script form in all commonly used computer platforms (Windows, Macintosh and Unix) during the past decade, there has been a tremendous information explosion of materials related to Tamil Cultural Heritage on the Net. Information exchange amongst Tamil Diaspora takes place via emails, mailing lists, websites of individuals and organizations (portal sites) using text based on 8-bit glyph-encoding. Project Madurai is one of several initiatives, aimed at electronic archiving of Tamil Literary works. Electronic texts of over 150 select Tamil literary works (ancient and contemporary) are available online through web-page delivery and via downloadable PDF files in TSCII (Tamil Script Code for Information Interchange, an 8-bit glyph-encoding).

Usage of 8-bit glyph-encoding is considered widely as interim measure, until the support for Unicode is available on commonly used server and desktop platforms. There is already concerted effort worldwide, particularly by major University libraries, to use Unicode to archive digital text material. Tamil has been fortunate amongst Indic languages to be implemented at the OS level in Windows and Linux during the past year. In addition, commercial databases like Oracle 8i and SQL Server support Unicode data in Tamil.

Text converters are now available to convert existing Tamil digital data in various 8-bit glyph-encoding to Unicode. Information exchange of Tamil content in Unicode is now possible through the Net via Web pages, PDF documents and various transfer-encoding formats used in electronic mail.

The goal of the present paper is to demonstrate the feasibility of a low-cost database solution for Unicode based Tamil e-texts, which allows search for availability of a specific work or specific word or word-string in a given electronic text. Prototype of a digital library will be demonstrated using select electronic texts of Tamil works in Unicode format, derived from the existing collections of Project Madurai. This can be readily extended to cover the vast amount of data already available at different Tamil Portal sites.

While there are commercial cataloguing software that employs standards recommended by the Library of Congress and support the use of Tamil text in Unicode, these can be inexpensive options for not-for-profit initiatives that are run entirely by volunteers.

The prototype will attempt to use a non-commercial database to store data and provide a browser based user interface for search and retrieval.