Migration need for Tamil users

N. Anbarasan
Chief Executive Officer
#39. 1st Cross, 1st Main, Shivanagar, W. C. Road, Bangalore - 560010
email : arasan@bgl.vsnl.net.in, aplesoft@vsnl.com

Abstract
Traditionally Tamil software users were acquainted with Tamil input interface software for their various needs. This has resulted into creation of documents in various file formats of the popular software like MS Word, MS PowerPoint, MS Excel, Pagemaker, Quark Express, Corel Draw etc. The Tamil software users are destined with Gigabytes of data, which includes proprietary encoding apart from the standard encoding adopted for Tamil.

As the international versions of shrink-wrapped packages like MS Office are more tightly integrated with standards, some of the codes already used in the standard encoding and proprietary encoding have already become unusable. This restricts the continuation of font hacking technology adopted for Tamil and forces the users to go for the universal standard called Unicode. This also ensures more stability of data created using Unicode.

The Unicode enabled applications support typing Tamil in Unicode only on the Unicode aware OS like MS Windows XP. In the absence of any alternate software for the Tamil users, the input interface software continues to play a rescuing role, still on the Unicode aware OS. Easy exchange of documents in Unicode can happen between MS Windows 98 and MS Windows XP users only when typing in Unicode on the ANSI based OS like MS Windows 98 is enabled. In the absence of any input interface for Tamil in Unicode on MS Windows 98, at least formatted document conversion which supports MS Word, MS Excel, MS PowerPoint, Pagemaker, Quark Express, Corel Draw etc is the only way out for Tamil users. This paper illustrates the needs for conversion tools or utilities and inevitable need for co-existence of Unicode aware applications on MS Windows 98.

Introduction

Lack of any international standards for Indian Languages to facilitate usage of Indian languages on the shrink-wrapped products of international version had forced the passionate language software developers to use hacked font encoding for Indian languages. No International software developer would include any language unless there is an international standard. In the absence of any international standards, restriction due to copyright and to respect the developer's IPR, the developers are forced to adopt their own proprietary encoding. When the international software giants left the Indian Languages for lack of standards and revenue, it is the local passionate developers who helped the language to sustain the technological onslaught. However, the hacked font techniques were misused by many developers and publishers to protect and promote their business interests. As a result there are tons of data generated using these proprietary encoding.

Even though, some Indian language data are available in the Internet, they are inaccessible to the general public due to their incompatible data format. When one intends to access any Indian language data on the Internet, the search engines fail to locate.

The non-standard hacked encoding also makes the documents created using such encoding as inaccessible forcing the users to go for retyping. This forces the users to convert the existing documents to standards before the situation becomes worst.

Indian language scenario

India is one of the pioneers in recognising the need for standards for scripts even before the formation of Unicode consortium. During 1983, Department of Electronics (DoE) formulated a standard called Indian Script Code for Information Interchange (ISCII). It was later revised during the year 1988 and was announced a standard by Indian standards institute Bureau of Indian Standards during 1991 as IS 13194: 1991.

This effort has enabled the Unicode consortium to quickly adopt the standard as part of Unicode standard.

Even though standards were in place, these were not implemented due to the limited market size. As a result, easy to develop hacked font techniques were popularly adopted by the Indian language software developers. Even after 20 years of realisation of need for standards, the basic technological requirement for the basic word processor has not been developed for Indian languages.

Font standards

On realizing the severity of the hacked encoding some of the state Governments have announced standards for their languages. As far as Tamil is concerned Tamilnadu state Government has announced two standards for Tamil to meet the requirements of Internet and publishing. In spite of the best efforts of the developers having adopted the standards, media centres of TV commercials and publishers continue to use their legacy systems.

Non-working of hacked font standards

As the International software developers are increasingly adopting the only International standard Unicode for Indian Languages, the days of hacked font techniques are coming to an abrupt end. This also implies that the software developed to input Indian languages text based on the hacked fonts for the shrink wrapped products like MS Office, are no more usable.

This author being a software developer for Indian Languages and had developed input/typing interface software for applications running on MS Windows, have come across “not-usable” codes in the usable ANSI codes. This problem is applicable even to the font standards like TAB, TAM and TSCII. In Powerpoint application of MS Office XP, the following codes are not usable:

167 - ‘U’ mathra used to write grantha uyimei letters.
168 - ‘UU’ mathra used to write grantha uyimei letters.
176 - Letter ‘ku’
Tamil Internet 2004, Singapore 3
177 - Letter ‘ngu’
180 - Letter ‘Tu’
215 - Letter ‘Voo’
247 - Letter ‘Lla’

This problem was even verified by the author of this paper with other input/typing interface software.

This author had already brought this to the notice of Microsoft, who is the developer of PowerPoint. Microsoft indicated that they couldn’t provide support for this problem. The only way out is to switch over to the International standard Unicode.

Stability of Unicode

Unlike the hacked font standards, which are pretending to be the standards for Indian languages, Unicode is protected by its stability policy and guarantees its usefulness in a meaningful way. As the hacked fonts are not recognised as standards by any standardisation body, the hacked font standards don’t have any sanctity from the standardisation institutions. Therefore, it is worth to switch over to Unicode.

Some of the shattered stumbling blocks

Some of the concerns of Indic computing community have been:

1. Unicode standard is based on Devanagari, which was adopted as base while encoding ISCII. Developers felt that Unicode (as it is based on the ISCII) do not satisfactorily Support non-Devanagari languages,

2. Encoding of character should be based on the linguistic nature of the language.

3. Character encoding order should meet the requirement of sorting.

Even though, Unicode has evolved from ISCII, Unicode has provision to improvise the encoding to meet the requirement, which is language specific and Unicode has considerably improved. Unfortunately, some developers block the adoption of Unicode with vested interests.

As the sorting order in a particular language is not really restricted to one order, it is evident that more than one sorter prevails even in Tamil. This nullifies the perception of the need for a character encoding to be based on the alphabetical order of a language. But still the requirement of Unicode to meet the needs of language specific writing system is yet to be addressed.

Usability of Unicode
It is obvious that any standards can be put into good use only when they are implemented. As the members of Unicode consortium are committed to implement Unicode in their products, more and more products would soon be available in the market, which is Unicode enabled.

At present, the availability of some application software for use in administration, DTP, education etc really encourages the user to go for Unicode based software. Some of the Unicode enabled software are MS Office XP, StarOffice 6.0, OpenOffice 1.1, CorelDraw 12, MS Internet Explorer 6.0 and Hot Potato 6.0

Some of the developer tools also support Unicode. Apart from this some more products, which are bound to adopt Unicode are Pagemaker, Photoshop, Quark Express, Word Perfect etc.

As the Unicode enabled applications are able to display Indic letters properly and meets the requirement of sorting along with the basic requirement of input, Unicode stands to prove that it deserves widest acceptance.

Enabling Windows 98 users to use Unicode
The installed user base of MS Windows 98 proves that MS Windows 98 still continues to be the dominating Operating System. For Unicode to be successfully adopted across the user segments amongst Unicode enabled Operating Systems and Unicode unaware OS, Unicode has to be enabled on the applications running on MS Windows 98. Otherwise, data/document portability across the platform would hinder the adaptability of Unicode.

Conversion of documents to Unicode
As the Unicode unaware Operating System like MS Windows 98 and Unicode aware Operating System and applications running on it will continue to co-exist till every user has upgraded their hardware, OS and the required applications to Unicode aware "ecosystem", there is a compelling need for conversion utilities to handle the documents created using hacked fonts and Unicode. Our past experience reminds that conversion of only text files or RTF files are of any (not) worth to try.

Needs of the users to migrate to the Unicode standards
Hacked font usage has penetrated to the extent that the users are not able to switchover even to the standard hacked fonts. It seems that a kind of fear looms large with the users to switchover which is coupled with the feelings that there is no imminent requirement for the switchover.

Users continue to use their systems including any legacy systems as long as they are able to use the system efficiently and it meets the requirement. However, when a need comes for the newer application software, it will not run on the older systems. A fact to be remembered here is that Operating Systems are hardware dependant. For example MS Windows 98 will not run on newer hardware with operating speed greater than 2.4 GHz. Similarly, the newer Application software also demands a version of Operating System as

a minimum requirement. This situation clearly overburdens the user when he thinks of any upgrade, which involves a financial dimension.

This author foresees a requirement for conversion tools, which can convert documents to and fro from hacked fonts to Unicode that would enable the users to utilise the available facilities effectively. With that in mind APPLESOFT has developed a suite of utilities to help the users of MS Office and Internet who use Internet Explorer.

Surabhi Tools

SURABHI TOOLS, is a suite of tools to support Indian Languages on MS Office running on MS Windows. This includes tools to cater for Sorting, Text conversion, Auto correct, Date and Time, Numerals to text etc.

Surabhi Tools retains the formatting layouts of the documents while supporting widely used hacked fonts and Unicode. Surabhi Tools frees the users from retyping or layout formatting and is an easy to use plug-in available within the application software such as MS Word, MS Excel, MS Access etc of MS Office.

Conclusion

Unicode is the only international standard for Indian Languages. Unicode is protected by its stability policies. And hence ensures usability of data generated in Unicode. Migration to Unicode is thus an inevitable one.