Challenges in implementation of Marathi Unicode

M. S. Sridhar

Joint Managing Director,
Cyberscape Multimedia Limited, Navi Mumbai.

Why do we need Unicode?

The Advantages

It has become mandatory to keep data in Unicode for Marathi as it is the only global standard followed across the globe. In this era of Information Technology, we are talking about global resource centers for data collection and dissemination. Unicode is supported by all the major Operating Systems. Unicode is being supported by all the major Operating Systems. Unicode is being supported by all new digital devices. With this background we can investigate the challenges in implementing

Unicode in Marathi.

- Unicode - Global Standard
- Unicode - Independent of Operating System
- Unicode - Supported by MS Office and Open Office
- Unicode - Supported by web browsers and web application development tools
- Unicode - Supported by Mobile Handsets for SMS
- Unicode - Supported by PDAs
- Unicode - Supported by search engines like Google

The Disadvantages

Unicode is a 16 bit coding system and hence the total data storage is double that of conventional non Unicode storage.

It does not work in older OS like Windows 98 etc.

It requires the OS to be loaded with the language components before using the same. The user has to configure the additional languages. Many users do not know this procedure or their version of OS does not have this facility. You require the OS that supports Marathi.

The Code Sequence

The table below gives the coding sequence for the Unicode page for Devagari that takes care of Hindi, Marathi and Sanskrit. The sequence covers the entire character set of Devnagari in the following manner.

Modifiers - chandra bindu, anuswar and visarga.

Vowels - aa to of.

Consonantas - ka, kha till ha.

Matras - Aa_matra to of_matra.

Vedic Characters - many symbols used in Vedic Sanskrit.

Nukta Characters - Ka, Kha, Ga, Ja, DA, DHa, Pha, Ya.

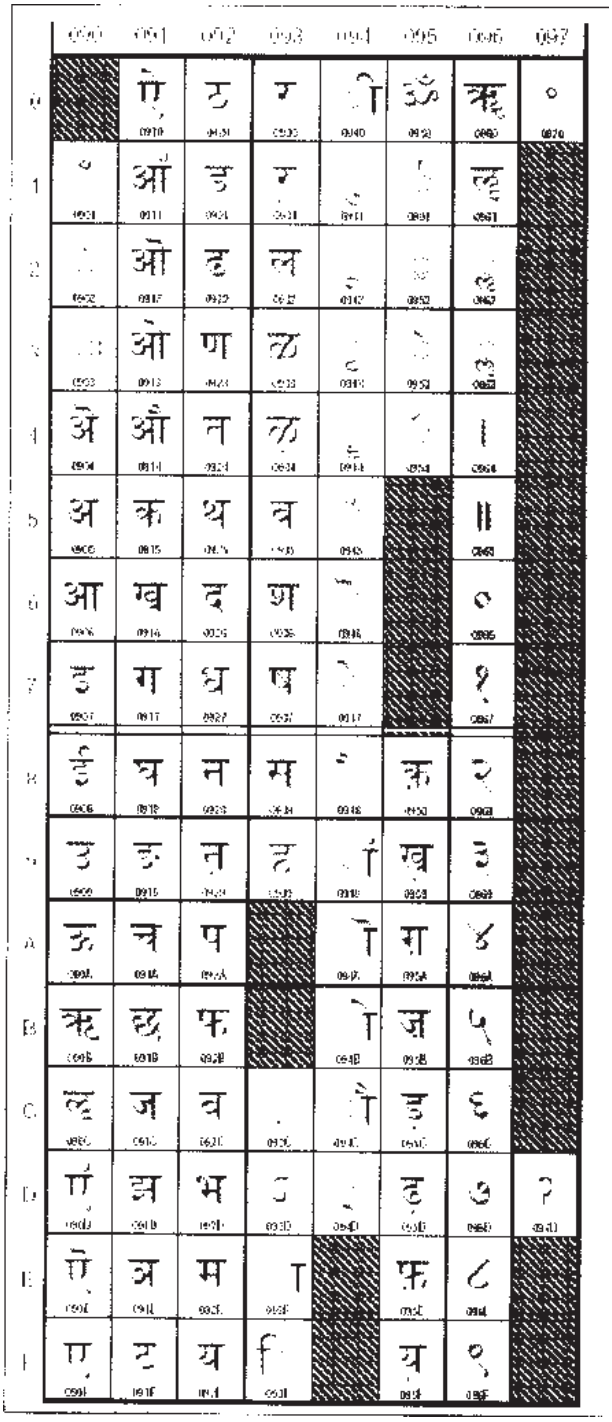
Numerals - 0 to 9 in Marathi.

Symbols - om, viram, double viram.

Marathi Characters - sha, la, Ra, Ya.

Sindhi Characters - GA, JA, DA, BA.

The code page spans from 0x0900 to 0x097D and has some blank unused characters in the middle to maintain consistency with other Indian Languages.



The ISCII Code sequence

The table below shows the character sequence in ISCII proposed in 1988 and later amended in 1991, the BIS standard for Indian Languages. The code ranges form 161 to 234. As you can see the Unicode character set has been adapted from ISCII and other symbols have have been included to accommodate the exceptions. Unicode has a direct one to one correlation with ISCII and it is very easy to convert the data to and from Unicode and ISCII.

ISCII Code	0	1	2	3	4	5	6	7
160		ँ	ं	ः	अ	आ	इ	ई
168	उ	ऊ	ऋ	ऐ	ए	ऐ	ऍ	ओ
176	ओ	औ	ऑ	क	ख	ग	घ	ङ
184	च	छ	ज	झ	ञ	ट	ठ	ड
192	ढ	ण	त	थ	द	ध	न	ऩ
200	प	फ	ब	भ	म	य	य़	र
208	ऱ	ल	ळ	ऴ	व	श	ष	स
216	ह	◌	ा	ि	ी	ु	ू	ृ
224	ॄ	ॅ	ॆ	े	ै	ॉ	ॊ	ो
232	ौ	्	।					

The Pre Requisites

Unicode is supported in Windows 2000, XP and 2003 and the new releases of Linux. We are restricting this discussion to Windows, as that is the most widely used Operating System. This indicates that Unicode cannot be effectively used in Windows 98 Operating System, that is most widely used in the state. Unicode implementation is very good in Windows XP coupled with MS Office XP, MS Office XP, MS Office 2003 and Open Office. Org version 1.4 onwards.

The Limiting Factors

Unicode is not usable in Windows 98. Windows 95 usage is minimal and can be ignored. It is partially implemented in Windows 2000 and Windows NT.

With such limitations, will we be able to used Unicode effectively ?

The solution and the path to follow

We have to realize that we cannot continue any longer following non-standard character coding and multiple character sets for Marathi. It is essential that we move to Unicode at the earliest. The action plan should be :

Convert all data to Unicode.

The new data should be generated in Unicode.

Systems having Windows 98 can have two way converters to convert Unicode data to Non Unicode data and vice versa.

This approach ensures that complete migration can be achieved in 2 to 3 years and we reap the benefits of moving to Unicode right from today.

The typical issues

In Marathi, we have two special characters that are extra. One is Ra and another is Ya. In Unicode 0930 is ra and 0931 is Ra which gets displayed as ra with a dot. In this position Marathi Ra can be used. For this purpose, we can generate a font that has that glyph in that position. Similarly 092F is ya and 095F us Ya. The character at 095F can be modified to accommodate the Marathi Ya.

There is a little difficulty in the collating sequence of the sorting order. The characters ksha, tra, dnya will not come after ha. They are conjuncts or jodakshars and will come in ka, ta and ja respectively, There is a facility to follow a special collating sequence in Ms SQL of Oracle database. This however will not work properly in other software and in The other characters that are typically different in Marathi are sha and la. These characters should be changed in the glyph level.

How to proceed with the conversions

The typical usage can be classified as

Word Processing - *.doc files.

Spreadsheets - *.mdb files.

Database - *.htm, *.html.

Text files - *.txt

The converters are capable of converting existing data in any of these formats any of the font to Unicode and vice versa. This will enable the data migration from legacy software.

The Unicode Fonts

The most popular Unicode fonts are Mangal in Widows, Raghu-8 Windows and Linux. There are may other Unicode fonts lime Marathi etc. The coding differences are there between the fonts. For that matter, the character Ta_Ta will look like this in ट्ट this in Raghu-8 and Qm ट्ट this in Mangal. Such differences exist in many characters. We need to understand that both are correct and simply two different forms of representing a combination. Today many aesthetically good looking typefaces are available for Marathi in Unicode. The user has a wide range to choose form and we have many vendors to provide the same.

The Unicode Web Sites

It is essential to develop and design web sites in Unicode. The main advantage to a Unicode enabled wed site is easy search. Search engines like Google etc. give search results in Unicode. It is possible to view these web sites in other Operating Systems also. With emerging technologies, It is possible to view Unicode data in a dynamic fashion with the help of dynamic fonts.

Unicode Web Application

We are at a stage where many applications are being developed over the Internet or Intranet. They are browser base applications. They can be from a imple interactive web site form to a vary large e-governance software that is deployed all over the state or country. It is essential such applications are developed in Unicode and the database is all maintained in Unicode and the database is also maintained in Unicode. Unicode database is very convenient to use

and it is very easy to offer facilities like indexing and sorting in Marathi. The web application requires a typing engine to type the data in Unicode.

The Tools

We at Cyberscape Multimedia Limited have developed the complete range of software tools to work in Unicode and to convert any type of data in any Marathi font to Unicode and vice versa. The tools can save invaluable time and provide and easy path to migrate to Unicode. We also have a special web enable that allows the user to display Unicode text and a facility to accept Unicode text and a facility to accept Unicode text from a Windows 98 computer. This will dramatically increase the usage of Unicode in the near future.

The conclusion

There are many hurdles before a complete migration is achieved in porting the data to Unicode. If we succumb to these hurdles, then we will have to face greater challenges later when Unicode will be very essential for any application. We advocate early use of these tools to migrate easily and effectively.

☆☆☆