

Unicode as a multilingual standard with reference to Indian languages

Rajesh Chandrakar

The author

Rajesh Chandrakar is Scientific and Technical Officer, INFLIBNET Centre, Ahmedabad, India.

Keywords

India, Languages, Archives, Data handling, Standards, Character recognition equipment

Abstract

India is a country rich in diversity in languages, cultures, customs and religions. Records of this complete culture, secret manuscripts and related documents of the respective religions, and 3,000 years of Indian history are available in their respective languages in different museums and libraries across the country. When the automation of libraries started in India, immediately the issue of localization of library and museum databases emerged. The issue became even more apparent with the advent of digital libraries and interoperability. At the start of automation, in the absence of proper standards, professionals tried to romanize documents as computers used to accept only binary digits of roman script to represent the English language. Later, the development of a new technology, ISCII, which is an extended form of the ASCII values from 126 to 255, helped library professionals in either developing the bilingual bibliographic databases or bilingual text files on DOS or Unix based applications. Gradually the font for Windows-based applications was developed for creating Web sites or document files. But now, with the requirement of different languages in the world including Indian, there is a forum available called "Unicode, Inc." which provides a solution to the localization problem of the world's languages. In this paper, Unicode as a multilingual standard is explained and the related technology available for localizing the Indian language materials is discussed.

Electronic access

The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0264-0473.htm

The Electronic Library
Volume 22 · Number 5 · 2004 · pp. 422-424
© Emerald Group Publishing Limited · ISSN 0264-0473
DOI 10.1108/02640470410561947

Introduction

Unicode is an encoding scheme designed to be a universal character set (UCS) for written characters and text; it is international in scope and includes characters from all of the major scripts of the world (more than 40,000). In addition, it includes commonly used technical symbols, ideographs, etc. using the same structure.

The Unicode consortium was incorporated in January 1991, under the name of Unicode, Inc. The Unicode provides a unique number for every character irrespective of platform, program, and the language[1]. Unicode is not hardware or software, it is a formal standard for multilingual encoding, which enables a single software product or a single Web site to be targeted across multiple platforms, languages and countries without re-engineering. It allows data to be transported through many different systems without corruption. Actually there are two separate efforts (one by Unicode, Inc. and another by ISO/IEC 10646) that have joined forces, yet remain separate entities. It is for this reason, that for general purposes mostly, the term "Unicodes" and "ISO/IEC 10646" are considered as synonymous.

Unicode was originally a simple, fixed-width 16-bit encoding. There was enough room in 16 bits for all modern writing systems under its initial design principles, but over the course of Unicode's growth and development, those principles had to give way. Eventually, 16 bits were no longer enough and needed an extension mechanism to allow for a longer number of characters. The standard mechanism uses pairs of Unicode values called surrogates to address over 1,000,000 possible values – around 1,114,112 characters total. As a result of different requirements, three different forms of Unicode – UTF8, UTF16 and UTF32 – have been introduced in the Unicode Standard, version 4.0.0[2]. In this version, 1,226 new character assignments have been added over and above what was in Unicode version 3.2.

Encoding schemes/forms of Unicode

The new edition of the Unicode, version 4.0.0, has tightened the definition of encoding terms, including UTF-32. Simple definitions of the encoding schemes are as follows:

- *UTF8*: attempts to allow legacy systems to use Unicode by coding the characters in the ASCII character set with only eight bits, and encoding characters that are not in the ASCII character set with 16 bits. This is commonly used for Web pages.
- *UTF16*: is a supplementary characters outside the basic multilingual plane. It encodes



maximum world's major languages in a fixed 16-bit character representation (2 bytes). This is the most common implementation.

- *UTF32*: is an actually UCS4 under a different name. It uses four bytes (32 bits) to encode all possible (millions) characters but is rarely used (Chandrakar, 2002).

To keep in view the nature of the Unicode character repertoire, versions of Unicode are defined in three ways, as follows[3]:

- (1) *Major*: the standard, which is published as a book.
- (2) *Minor*: minor version covers the character additions or more significant normative changes, which is published as a technical report.
- (3) *Update*: any other changes to normative or important informative portions of the standard that could change program's behavior. These changes are reflected in new Unicode character database files and update pages.

Apart from this list, the major version or edition of the Unicode, is published in book form such as Unicode version 1.0.0, published in 1991. So far Unicode version 3.0.0 is called the major edition of the Unicode and was published in 2000. The latest and major edition of the Unicode is version 4.0.0 introduced in August 2003 (Unicode Consortium, 2003), although a beta version of 4.0.1 has since been released for public review.

Indian languages and Unicode

According to the Unicode version 3.0.0, Indian languages have been allocated place in chapter 9 ("South and Southeast Asian scripts") covers almost 15 different scripts including nine Indian ones. The Urdu language has been allocated a separate place in chapter 8 ("Middle Eastern scripts") along with Hebrew, Syriac and Thana scripts. The Unicode Consortium has allocated the unique numbers to the scripts shown in Table I.

Table I The Unique numbers to the scripts allocated by the Unicode Consortium

Script	Assigned unique number by Consortium
Arabic	U+0600-U+06FF (01536-01791)
Devnagari	U+0900-U+097F (02304-02431)
Bengali	U+0980-U+09FF (02432-02559)
Gurumukhi	U+0A00-U+0A7F (02560-02687)
Gujarati	U+0A80-U+0AFF (02688-02815)
Oriya	U+0B00-U+0B7F (02816-02943)
Tamil	U+0B80-U+0BFF (02944-03071)
Telugu	U+0C00-U+0C7F (03072-03199)
Kannada	U+0C80-U+0CFF (03200-03327)
Malayalam	U+0D00-U+0D7F (03328-03455)

In addition to the above list of Indian scripts and their characters range, many individual new characters are added in to the latest edition of the standard (i.e. version 4.0.0) such as currency symbols including substantial improvements to the script descriptions.

Available multi-script fonts for Indian languages

For creating Web pages, some individual institutes such as commercial, newspaper agencies, and business institutes have taken initiatives with the help of some true type font (TTF). For instance, www.webdunia.com works on [webdunia.ttf](http://www.webdunia.ttf) and www.bhaskar.com works on [bhaskar.ttf](http://www.bhaskar.ttf) etc. To browse the information hosted on those Web sites requires installing the appropriate font at the client side. Nowadays, some institutes are using dynamic fonts for creating Web pages, which does not required to install fonts at the client side. This kind of solution is perfect for Web page creation, but as far as databases are concerned, this is not a proper solution of the problem and cannot be implemented as well. To be frank, with the help of either ISCII based products or ISFOC, or any other true type font, one can develop the bi-script database of Indian languages with a combination of roman script successfully. But a problem arises either when someone wants to have multi-script data records in a database or when one wants to share their data with other systems or create either multilingual or multi-script electronic resources. As far as Web sites are concerned, Unicode-based Indian multilingual language processing can be seen on the Indian Google Web site (www.google.co.in), which is dedicated to some of the Indian languages such as Hindi, Bengali, Telugu, Marathi, and Tamil. But, the Unicode Web site itself (www.unicode.org) is an excellent example of global multilingual processing.

Some localization industries has developed different Unicode-based fonts, which are available on the market – some of them are free and some of them not. But most of the fonts available are for the individual/single scripts such as "mangal" for Devnagari script, etc. As for as the multi-script font for Indian languages is concerned, there are only two fonts available which covers almost all the Indian languages: i.e. Arial Unicode MS and Code2000. Both are Unicode-based open type fonts. Arial Unicode MS always comes with the Windows 2000 and higher operating systems developed by Microsoft, while Code2000 is developed by James Kass. A demo copy of the font is available on his personal Web site: <http://home.att.net/~jameskass/>

Requirements for Indian language processing (ILP)

Bearing in mind India as a large multi-lingual society with as many as eighteen constitutionally recognized languages including English and the National language Hindi, the development of an Indian Language Processing (ILP) system could be designed using many approaches such as[4]:

- national language interface/environment for data input/output support;
- operating system level support at the native level for the Indian languages;
- Indian language shell over the existing operating systems and applications;
- localizing existing applications;
- developing specific applications; and
- designing language compilers in natural languages.

Out of those mentioned approaches, the second one is appropriate and easy to implement, if one is going to develop the new software, as it does not require either special staff training or special keyboards or fonts for input and retrieve data.

To discuss the implementation problem of the Indian languages on Unicode, a separate user forum has been created by Unicode. Queries can be sent to indic@unicode.org

Conclusion

As far as the encoding of Indian languages are concerned the Unicode Consortium came out with a solution during 2000 with its 3.0.0 version of the Unicode standard and the shortcomings of

this version have been addressed in the latest version 4.0.0. Based on the above standard, the MARC Development Group at the Library of Congress has recently come out with a report on the assessment of options for handling full Unicode character encoding in MARC21 for multilingual databases, which also includes Indian languages[5]. Basically this encoding scheme is a replacement of the MARC8 Encoding Scheme, where each character is encoded using a single 8-bit byte unlike Unicode which is multi-byte per character code set. In so far as Indian languages are concerned, the Unicode-based product is the only solution for new terminology.

Web sites

- 1 Unicode, Inc.: www.unicode.org
- 2 Unicode 4.0.0: www.unicode.org/versions/Unicode4.0.0/
- 3 Versions of the Unicode Standard: www.unicode.org/standard/versions/
- 4 Introduction to TDIL Programme: <http://tdil.mit.gov.in/introindx.html>
- 5 Assessment of options for handling full Unicode character encoding in MARC21: www.loc.gov/marc/marbi/2004/2004-report01.pdf

References

- Chandrakar, R. (2002), "Multi-script bibliographic database: an Indian perspective", *Online Information Review*, Vol. 26 No. 4, pp. 246-51.
- Unicode Consortium (2003), "The Unicode Standard, version 4.0.0", defined by: *The Unicode Standard, Version 4.0*, Addison-Wesley, Reading, MA.

About the author

Rajesh Chandrakar holds a Bachelor's degree in Science (Physics, Chemistry and Mathematics) from the Government Model College of Science, Raipur, Chhattisgarh, India, and a Master's degree in Library and Information Science from Pt. Ravishankar Shukla University, Raipur, Chhattisgarh, India. He also holds a Postgraduate diploma in Computer Applications from the same university. He is currently Scientific and Technical Officer at the INFLIBNET (Information and Library Network) Centre, Ahmedabad (Gujarat), India where he has been working on developing union databases on different core resources of library such as books, serials, theses, etc. for the past seven years. In addition, he is involved with the software development team and is also working with SOUL (Software for University Library) software. He has also been assigned to the activities related to cataloguing and bibliographic standards at the Centre. In this regard, he is acting as Convenor of the MARC21 Core Group of INFLIBNET Centre including the alternative member of Bureau of Indian Standards (BIS), Technical Committee MSD 5. He can be contacted at: Rajesh Chandrakar (STO-I), INFLIBNET Centre, UGC, P.B. No. 4116, Navrangpura, Ahmedabad 380 009, Gujarat, India. Tel: +91-(0)79-6305971, 6304695,6308528; Fax: +91-(0)79-6300990, 6397816; E-mail: rajesh@inflibnet.ac.in or rchandrakar@yahoo.com