

The Unique Unicode

by Mir Lutful Kabir Saadi

[Mir Lutful Kabir Saadi <mirsaaadi@sdbnd.org> is a Fellow, 21st Century Trust, UK, and Dhaka Bureau Chief, IMPACT International, UK., and General Secretary, Bangladesh Science Writers & Journalists Forum]

The Unicode character set is a character set projected to represent the writing schemes of all of the world's major languages. Unicode provides a unique number for every character, no matter what the platform, what the programme, or what the language.

The Unicode Standard was created by a team of computer professionals, linguists, and scholars to become a worldwide character standard, one easily used for text encoding everywhere. To that end, the Unicode Standard follows a set of fundamental principles: Universal repertoire, logical order, efficiency, unification, characters not glyphs, dynamic composition, semantics, equivalent sequence, plain text and convertibility.

The Unicode Standard is a character coding system designed to support the worldwide interchange, processing, and display of the written texts of the diverse languages of the modern world. In addition, it supports classical and historical texts of many written languages. Formally, the Unicode standard is defined by the latest printed version of the book The Unicode Standard, plus online documents and data that update and extend the book's normative specifications and informative content.

The Unicode is developed by the Unicode Consortium. The Unicode Consortium is a non-profit organisation founded to develop, extend and promote use of the Unicode Standard, which specifies the representation of text in modern software products and standards. The membership of the consortium represents a broad spectrum of corporations and organisations in the computer and information processing industry. The consortium is supported financially solely through membership dues. Membership in the Unicode Consortium is open to organisations and individuals anywhere in the world who support the Unicode Standard and wish to assist in its extension and implementation.

The Unicode Standard defines codes for characters used in all the major languages written today. Scripts include the European alphabetic scripts, Middle Eastern right-to-left scripts, and many scripts of Asia. The Unicode Standard further includes punctuation marks, diacritics, mathematical symbols, technical symbols, arrows, dingbats, etc. It provides codes for diacritics, which are modifying character marks such as the tilde (~), that are used in conjunction with base characters to represent accented letters (ñ, for example). In all, the Unicode Standard, Version 3.2 provides codes for 95,221 characters from the world's alphabets, ideograph sets, and symbol collections.

Basically, computers just deal with numbers. They store letters and other characters by conveying a number for each one. Before Unicode was invented, there were hundreds of different encoding systems for assigning these numbers. No single encoding could contain enough characters. For example, the European Union alone requires several different encodings to cover all its languages. Even for a single language like English no single encoding was adequate for all the letters, punctuation and technical symbols in common use.

These encoding systems also conflict with one another. That is, two encodings can use the same number for two different characters, or use different numbers for the same character. Any given computer (especially servers) needs to support many different encodings; yet whenever data is passed between different encodings or platforms, that data always runs the risk of corruption.

The Unicode Standard has been adopted by such industry leaders as Apple, HP, IBM, Just System, Microsoft, Oracle, SAP, Sun, Sybase, Unisys and many others. Unicode is required by modern standards such as XML, Java, ECMA Script (JavaScript), LDAP, CORBA 3.0, WML, etc., and is the official way to implement ISO/IEC 10646. It is supported in many operating systems, all modern browsers, and many other products. The emergence of the Unicode Standard, and the availability of tools supporting it, is among the most significant recent global software technology trends.

Incorporating Unicode into client-server or multi-tiered applications and websites offers significant cost savings over the use of legacy character sets. Unicode enables a single software product or a single website to be targeted across multiple platforms, languages and countries without re-engineering. It allows data to be transported through many different systems without corruption.

The primary feature of Unicode 3.2 is the addition of 1016 new encoded characters. These additions consist of several Philippine scripts, a large collection of mathematical symbols, and small sets of other letters and symbols. All of the newly encoded characters in Unicode 3.2 are additions to the Basic Multilingual Plane (BMP). Unicode 3.2 also features amended contributory data files, to bring the data files up to date against the expanded repertoire of characters. All outstanding errata and corrigenda to the Unicode Standard are included in this specification.

The design of Unicode is based on the simplicity and consistency of ASCII, but goes far beyond ASCII's limited ability to encode only the Latin alphabet. The Unicode Standard provides the capacity to encode all of the characters used for the written languages of the world. To keep character coding simple and efficient, the Unicode Standard assigns each character a unique numeric value and name.

The Unicode Standard directly addresses only the encoding and semantics of text. It addresses no other action performed on the text. For example, the word processor may check the typist's input as it is being entered, and display misspellings with a wavy underline. Or it may insert line breaks when it counts a certain number of characters entered since the last line break. An important principle of the Unicode Standard is that it does not specify how to carry out these processes as long as the character encoding and decoding is performed properly.

The Unicode Standard specifies an algorithm for the presentation of text with bidirectional behaviour, for example, Arabic and English. Characters are stored in logical order. The Unicode Standard includes characters to specify changes in direction when scripts of different directionality are mixed. For all scripts Unicode text is in logical order within the memory representation, corresponding to the order in which text is typed on the keyboard.

You may not find the character in what you think is the obvious spot. While the characters in Unicode are grouped into blocks, this is only a rough grouping because

characters can be categorised many different ways. In particular, punctuations and symbols are applicable across a very wide range of usages and scripts (writing systems). Even the notion of a script itself is not well-defined; text in a given language may make use of characters from multiple scripts.

Finally, your character may not yet be encoded in Unicode. There is a well defined submission process for new characters or scripts. This process verifies that the proposed character is in fact a candidate for encoding. In some cases, this process may not be straightforward: for example, Egyptian hieroglyphs have not yet been encoded because there is not yet general agreement on the exact repertoire of characters.

[End]

Source: Ubiquity Volume 6, Issue 21 (June 15-22, 2005)
<http://www.acm.org>