Chapter XVII

Unicode Characters for Human Dentition: New Foundation for Standardized Data Exchange and Notation in Countries Employing Double-Byte Character Sets

Hiroo Tamagawa
The Japan Association for Medical Informatics, Japan

Hideaki Amano
The Japan Association for Medical Informatics, Japan

Naoji Hayashi
The Japan Association for Medical Informatics, Japan

Yasuyuki Hirose
The Japan Association for Medical Informatics, Japan

Masatoshi Hitaka
The Japan Association for Medical Informatics, Japan

Noriaki Morimoto
The Japan Association for Medical Informatics, Japan

Hideaki Narusawa
The Japan Association for Medical Informatics, Japan

Ichiro Suzuki
The Japan Association for Medical Informatics, Japan

ABSTRACT

In this chapter, the authors report the minimal set of characters from the Unicode Standard that is sufficient for the notation of human dentition in Zsigmondy-Palmer style. For domestic reasons, the Japanese Ministry of International Trade and Industry expanded and revised the Japan Industrial Standard (JIS) character code set in 2004 (JIS X 0213). More than 11,000 characters that seemed to be necessary for...
Unicode Characters for Human Dentition
denoting and exchanging information about personal names and toponyms were added to this revision, which also contained the characters needed for denoting human dentition (dental notation). The Unicode Standard has been adopted for these characters as part of the double-byte character standard, which enabled, mainly in eastern Asian countries, the retrieval of human dentition directly on paper or displays of computers running Unicode-compliant OS. These countries have been using the Zsigmondy-Palmer style of denoting dental records on paper forms for a long time. The authors describe the background and the application of the characters for human dentition to the exchange, storage and reuse of the history of dental diseases via e-mail and other means of electronic communication.

INTRODUCTION

Computers store letters and other characters by assigning a number to each character (Kilbourne and Williams, 2003). The entire collection of characters is called a character code set. The character code set was established in Europe and America, and was gradually expanded together with the advent of computers. ASCII (American Standard Code for Information Interchange) (ANSI INCITS 4-1986, 1963) is one of the most popular and representative character code sets, and it is capable of encoding a maximum of 256 characters using one byte of information. All alphanumeric characters are covered by this code system, and ASCII has become the de facto worldwide standard due to its simplicity.

Together with the development of computers, a problematical point emerged in that the one-byte character set does not have the capacity to encode enough characters in countries such as Japan, China, Korea and other Asian countries which utilize ideographic writing systems (Hussein et al. 2004). The majority of these countries employs writing systems based on the so-called Chinese characters and has their origins in long-standing Asian culture. Other glyphs, symbols and icons developed in the course of history of the individual countries are also used.

In order to circumvent the limitations of the ASCII character set, computer systems in these countries use two bytes for representing each character. Characters encoded using two-byte codes are called double-byte characters (Oram 1991). In double-byte character sets, each character is represented by two bytes, which enables the encoding of a maximum of 65536 (256x256) characters.

The design of Unicode (http://www.unicode.org/) is based on the simplicity and the consistency of ASCII; however, since it assigns between one and four bytes to each character, it goes far beyond the abilities of ASCII which is limited to encoding the Latin alphabet and a number of other characters, including those with diacritic marks used in European languages. Before Unicode was invented, there were hundreds of different encoding systems for assigning these numbers as single encoding could contain enough characters (http://en.wikipedia.org/wiki/Unicode). For example, the European Union alone requires several different encodings to cover all its languages and even for a single language like English, no single encoding was adequate for covering all the letters, punctuation, and technical symbols in common use.

Unicode has the explicit aim of transcending the limitations of traditional character encodings, such as those defined by the ISO 8859 standard (ISO/IEC 8859-11, 1999), which are widely used in various countries of the world but remain largely incompatible with each other. In principle, Unicode encodes the underlying characters, graphemes and grapheme-like units rather than