Appendix C. Unicode UTF-8 encoding

UTF-8 is an efficient encoding of Unicode character - String that recognizes the fact that the majority of text-based communications are in ASCII. It therefore optimizes the encoding of these characters.

Unicode is preferred to ASCII because it permits the inclusion of accents, scientific symbols and characters used in languages other than English. The UTF-8 format is a standard encoding that provides the most efficient means of encoding 16-bit Unicode characters in cases where the majority of characters are in the ASCII range. Both UTF-8 and the alternative UTF-16 encoding are supported by all widely used operating systems and major applications. UTF-8 was adopted is an IETF Internet Standard (it was initially adopted by IETF in 1996 to restrict some code values in 1998 and 2003). In 2008 UTF-8 became the most widely used for of encoding in web pages.

SNOMED CT uses the UTF-8 representation[1] of characters in terms and other text fields.

Footnotes

RefNotes

[1] Note that SNOMED CT does not use, or require use of, the Byte Order Mark (BOM) specified by the Unicode standard because all SNOMED CT release files use UTF-8.