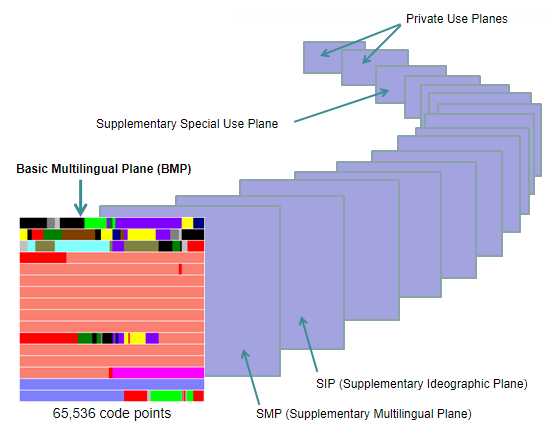
Character Encoding Algorithms

Unicode

It is sort of like a universal character set. In its character set the first 65,536 code point positions are said to constitute the BMP (Basic Multilingual Plane). BMP includes the most commonly used characters. The Unicode character set also contains space for around a million additional code point positions. Characters in this latter range are referred to as supplementary characters.



Unicode has multiple ways of encoding like UTF-8, UTF-16, UTF-32

UTF-8 uses 1 byte to represent characters in the ASCII set, two bytes for characters in several more alphabetic blocks, and three bytes for the rest of the BMP. Supplementary characters use 4 bytes.

UTF-16 uses 2 bytes for any character in the BMP, and 4 bytes for supplementary characters.

UTF-32 uses 4 bytes for all characters.

