

❖ What is ASCII and UTF-8?

➤ ASCII:

ASCII: abbreviated from **American Standard Code for Information Interchange**, is a character encoding standard for electronic communication. ASCII codes represent text in computers, telecommunications equipment, and other devices. Most modern character-encoding schemes are based on ASCII, although they support many additional characters.

ASCII is the traditional name for the encoding system; the Internet Assigned Numbers Authority (IANA) prefers the updated name **US-ASCII**, which clarifies that this system was developed in the US and based on the typographical symbols predominantly in use there.

Originally based on the English alphabet, ASCII encodes 128 specified characters into seven-bit integers. Ninety-five of the encoded characters are printable: these include the digits 0 to 9, lowercase letters *a* to *z*, uppercase letters *A* to *Z*, and punctuation symbols. In addition, the original ASCII specification included 33 non-printing control codes which originated with Teletype machines; most of these are now obsolete, although a few are still commonly used, such as the carriage return, line feed and tab codes.

➤ UTF-8:

UTF-8 (Unicode Transformation Format-8 bit): is a variable width character encoding capable of encoding all 1,112,064 valid code points in Unicode using one to four 8-bit bytes. The encoding is defined by the Unicode standard, and was originally designed by Ken Thompson and Rob Pike. The name is derived from *Unicode* (or *Universal Coded Character Set*) *Transformation Format – 8-bit*.

UTF-8 is a standard for representing Unicode numbers in computer files. Symbols with a Unicode number from 0 to 127 are represented exactly the same as in ASCII, using one 8-bit byte. This includes all Latin alphabet letters without accents.

It was designed for backward compatibility with ASCII. Code points with lower numerical values, which tend to occur more frequently, are encoded using fewer bytes. The first 128 characters of Unicode, which correspond one-to-one with ASCII, are encoded using a single octet with the same binary value as ASCII, so

that valid ASCII text is valid UTF-8-encoded Unicode as well. Since ASCII bytes do not occur when encoding non-ASCII code points into UTF-8, UTF-8 is safe to use within most programming and document languages that interpret certain ASCII characters in a special way.