HTML Unicode (UTF-8) Reference

The Unicode Character Sets

Unicode can be implemented by different character sets. The most commonly used encodings are UTF-8 and UTF-16:

|  |  |
| --- | --- |
| **Character-set** | **Description** |
| UTF-8 | A character in UTF8 can be from 1 to 4 bytes long. UTF-8 can represent any character in the Unicode standard. UTF-8 is backwards compatible with ASCII. UTF-8 is the preferred encoding for e-mail and web pages |
| UTF-16 | 16-bit Unicode Transformation Format is a variable-length character encoding for Unicode, capable of encoding the entire Unicode repertoire. UTF-16 is used in major operating systems and environments, like Microsoft Windows, Java and .NET. |

**Tip:** 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, making valid ASCII text valid UTF-8-encoded Unicode as well.

HTML 4 supports UTF-8. HTML 5 supports both UTF-8 and UTF-16!

## **The Difference Between Unicode and UTF-8**

Unicode is a **character set**. UTF-8 is **encoding**.

Unicode is a list of characters with unique decimal numbers (code points). A = 65, B = 66, C = 67, ....

This list of decimal numbers represent the string "hello": 104 101 108 108 111

Encoding is how these numbers are translated into binary numbers to be stored in a computer:

UTF-8 encoding will store "hello" like this (binary): 01101000 01100101 01101100 01101100  01101111

**Encoding** translates numbers into binary. **Character sets** translates characters to numbers.