HTML - Character Encodings

Character encoding is a method of converting bytes into characters. To validate or display an HTML document properly, a program must choose a proper character encoding.

HTML Charset Attribute

The HTML charset attribute of meta tag is used to mention character encoding of webpage.

```
<meta charset="UTF-8">
```

The ASCII Character Set

The most common character set or character encoding in use on computers is ASCII **(The American Standard Code for Information Interchange)**, and this is probably the most widely used character set for encoding text electronically. ASCII encoding consist of 128 characters(0-127).

- English Alphabets (A-Z and a-z)
- Numbers(0-9)
- Special Characters (@, #, $, %, etc)

You can have a look at complete set of **Printable ASCII Characters**

The ANSI Character Set

ANSI character set is generally used in windows systems, it is also called as windows-1252. This includes

- From 0 to 127 ANSI follows ASCII characters.
- From 128 to 159 some extra special characters are added.
- From 160 to 255 it's identical to UTF-8.

The ISO-8859-1 Character Set

ISO-8859-1 was the default character set for HTML 4. This character set supported 256 different character codes.

- Same as ASCII for the first 128 characters
- Does not use the characters from 128 to 159

- Same as ANSI and UTF-8 from 160 to 255

The UTF-8 Character Set

The HTML5 specifications recommends developers to use UTF-8 encodings in webpages, because UTF-8 covers all character and symbols in the world. The characters of UTF-8 are.

- Identical to ASCII for 0 to 127 characters
- Characters 128 to 159 are empty
- Uses same characters as ANSI and 8859-1 from 160 to 255
- Characters from other language are specified using 256 to 1000

The International Standards Organization created a range of character sets to deal with different national characters. For the documents in English and most other Western European languages, the widely supported encoding ISO-8859-1 is used.

ISO Character Sets

Here is the list of Character Set being used around the world along with their description.

Character Set	Description
ISO-8859-1	Latin alphabet part 1 Covering North America,Western Europe, Latin America, theCaribbean, Canada, Africa
ISO-8859-2	Latin alphabet part 2 Covering Eastern Europe
ISO-8859-3	Latin alphabet part 3 Covering SE Europe, Esperanto, miscellaneous others
ISO-8859-4	Latin alphabet part 4 Covering Scandinavia/Baltics (and others not in ISO-8859-1)
ISO-8859-5	Latin/Cyrillic alphabet part 5
ISO-8859-6	Latin/Arabic alphabet part 6
ISO-8859-7	Latin/Greek alphabet part 7
ISO-8859-8	Latin/Hebrew alphabet part 8
ISO-8859-9	Latin 5 alphabet part 9 Same as ISO-8859-1 except Turkish characters replace Icelandic ones
ISO-8859-10	Latin 6 Latin 6 Lappish, Nordic, and Eskimo
ISO-8859-15	The same as ISO-8859-1 but with more characters added

ISO-2022-JP	Latin/Japanese alphabet part 1
ISO-2022-JP-2	Latin/Japanese alphabet part 2
ISO-2022-KR	Latin/Korean alphabet part 1

The Unicode Consortium was then set up to devise a way to show all characters of different languages, rather than have these different incompatible character codes for different languages.

Therefore, if you want to create documents that use characters from multiple character sets, you will be able to do so using the single Unicode character encodings.

Unicode therefore specifies encodings that can deal with a string in special ways so as to make enough space for the huge character set it encompasses. These are known as UTF8, UTF-16, and UTF-32.

UTF Character Sets

Character Set	Description
UTF-8	A Unicode Translation Format that comes in 8-bit units that is, it comes in bytes. A character in UTF8 can be from 1 to 4 bytes long, making UTF8 variable width.
UTF-16	A Unicode Translation Format that comes in 16-bit units that is, it comes in shorts. It can be 1 or 2 shorts long, making UTF16 variable width.
UTF-32	A Unicode Translation Format that comes in 32-bit units that is, it comes in longs. It is a fixed-width format and is always 1 "long" in length.

The first 256 characters of Unicode character sets correspond to the 256 characters of ISO-8859-1. By default, HTML 4 processors should support UTF-8, and XML processors are supposed to support UTF-8 and UTF-16; therefore all XHTML-compliant processors should also support UTF-16.