Print Page

HTML Character Sets

« Previous (ref_colormixer.asp Next Reference » (ref_urlencode.asp

To display an HTML page correctly, the browser must know what character set (character encoding) to use.

HTML Character Sets

What is the correct character encoding to use in HTML?

For HTML5, the default character encoding is UTF-8.

This has not always been the case. The character encoding for the early web was ASCII.

Later, from HTML 2.0 to HTML 4.01, ISO-8859-1 was considered the standard.

With XML and HTML5, UTF-8 finally arrived, and solved a lot of character encoding problems.

Below is a brief description of the character encoding standards.

In the Beginning: ASCII

Computer information (numbers, texts, and pictures) is stored as binary ones and zeros (01000101) in the electronics.

To standardize the storing of alphanumeric characters, the American Standard Code for Information Interchange (ASCII) was created. It defined a unique binary 7-bits number for each storable character to support the numbers from 0-9, the upper/lower case English alphabet (a-z, A-Z), and some special characters like ! $ + - ( ) @ < > .

Since ASCII used one byte (7 bits for the character, and one of bit for transmission parity control), it could only represent 128 different characters. In addition 32 of these characters were reserved for other control purposes.

The biggest weakness with ASCII was that it excluded non English letters.

ASCII is still in widespread use today, especially in large mainframe computer systems.

For a closer look, please study our Complete ASCII Reference (../charsets/ref_html_ascii.asp).

In Windows: ANSI

ANSI (also called Windows-1252) was the default character set in Windows, up to Windows 95.

ANSI is an extension to ASCII, with added international characters. It uses a full byte (8-bits) to represent 256 different characters.

Since ANSI has been the default character set in Windows, it is supported by all browsers.

For a closer look, please study our Complete ANSI Reference (../charsets/ref_html_ansi.asp).

In HTML 4: ISO-8859-1

Since most countries use characters outside ASCII, the default character encoding in the HTML 2.0 standard was changed to ISO-8859-1.

ISO-8859-1 is an extension to ASCII, with added international characters. Like ANSI, it uses a full byte to represent twice as many characters than ASCII.

When browsers detect ISO-8859-1 in a web page, they normally default to ANSI, because ANSI is identical to ISO-8859-1 except that ANSI has 32 extra characters.

If an HTML 4 web page uses a different character-set than ISO-8859-1, it should be specified in the <meta> tag:

Example

```
<meta http-equiv="Content-Type" content="text/html;charset=ISO-8859-8">
```

Try it Yourself » (tryit.asp?filename=tryhtml_charsets)

The default character set for HTML5 is UTF-8.
All HTML 4 processors support UTF-8, and all HTML5 and XML processors support both UTF-8 and UTF-16.

For a closer look, please study our Complete ISO-8859-1 Reference (../charsets/ref_html_8859.asp).

In HTML5: Unicode UTF-8

Because the character sets listed above are limited, and not compatible in multilingual environments, the Unicode Consortium developed the Unicode Standard.

The Unicode Standard covers (almost) all the characters, punctuations, and symbols in the world.

Unicode enables processing, storage, and transport of text, independent of platform and language.

The default character encoding in HTML5 is UTF-8.

For a closer look, please study our Complete Unicode Reference (../charsets/ref_html_utf8.asp).

« Previous (ref_colormixer.asp Next Reference » (ref_urlencode.asp