HTML5 - CHARACTER ENCODINGS

A character encoding is a method of converting bytes into characters. To validate or display an HTML document, a program must choose a character encoding. HTML 5 authors have three means of setting the character encoding −

HTTP Content-Type Header

If you are writing cgi or similar program then you would use HTTP *Content-Type* header to set any character encoding.

Following is the simple example −

```
print "Content-Type: text/html; charset=utf-8\r\n";
```

The <meta> element

You can use a <meta> element with a charset attribute that specifies the encoding within the first 512 bytes of the HTML5 document.

Following is the simplified example −

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

Above syntax replaces the need for <meta http-equiv="Content-Type" content="text/html; charset=UTF-8"> although that syntax is still allowed.

Unicode Byte Order Mark *BOM*

A byte order mark *BOM* consists of the character code U+FEFF at the beginning of a data stream, where it can be used as a signature defining the byte order and encoding form, primarily of unmarked plaintext files.

Many Windows programs *includingWindowsNotepad* add the bytes 0xEF, 0xBB, 0xBF at the start of any document saved as UTF-8. This is the UTF-8 encoding of the Unicode byte order mark *BOM*, and is commonly referred to as a UTF-8 BOM even though it is not relevant to byte order.

For HTML5 document, you can use a Unicode Byte Order Mark *BOM* character at the start of the file. This character provides a signature for the encoding used.