

## Re: Acquiring UTF-8 string length

---

*Source:* <http://www.tech-archive.net/Archive/VC/microsoft.public.vc.language/2007-04/msg00017.html>

---

- *From:* Ulrich Eckhardt <[eckhardt@xxxxxxxxxxxxxxxx](mailto:eckhardt@xxxxxxxxxxxxxxxx)>
  - *Date:* Mon, 02 Apr 2007 09:39:25 +0200
- 

Coder Guy wrote:

Well as in the example at [http://en.wikipedia.org/wiki/Multi-byte\\_character\\_set](http://en.wikipedia.org/wiki/Multi-byte_character_set), this UTF-8 string actually has three different lengths:

```
// I{heart}NY
char str[] = { 0x49, 0xE2, 0x99, 0xA5, 0x4E, 0x59, 0x00 };
```

count of bytes = 7, obtained by sizeof()

Right.

count of code points = 6, obtained by strlen()

Wrong. strlen() only returns the number of chars up to the first NUL char. The number of codepoints is four, plus the terminating NUL.

count of characters = 4, no API as far as I can tell

Four characters plus a terminating NUL. Note that there are codepoint that don't resolve to a character and characters that resolve to more than one codepoint.

How would I get the number of characters in this string? Or how would I go about reversing the characters in this string? Do I have to really implement my own UTF-8 decoder/encoder?

There are libraries out there that help you handle all the various facets of Unicode. You might want to take a look at ICU, for example.

Re: Acquiring UTF-8 string length

Another example might be that I am reading a file and it specifies a code page at the top... aren't there any APIs which will help me manage by per-character rather than per-code point?

Not that I was aware of.

[snipped fullquote with signature]

Stop this misbehaviour please, it is considered impolite on the Usenet.

Uli

.