What's the difference between Text Document, Text Document – MS-DOS Format, and Unicode Text Document?

devblogs.microsoft.com/oldnewthing/20120220-00

February 20, 2012

Raymond Chen

Alasdair King asks why Wordpad has three formats, *Text Document*, *Text Document – MS-DOS Format*, and *Unicode Text Document*. "Isn't at least one redundant?" Recall that in Windows, three code pages have special status.

1. Unicode (more specifically, UTF-16LE)
2. `CP_ACP`, commonly known as the ANSI code page, although that is a misnomer
3. `CP_OEM`, commonly known as the OEM code page, although that too is a misnomer.

Three text file formats. Three encodings. Hm… I wonder… As you might have guessed by now, the three text file formats correspond to the three special code pages. Now it's just a matter of deciding which one matches with which. The easiest one is the Unicode one; it seems clear that *Unicode Text Document* matches with Unicode. Okay, we now have to figure out how *Text Document* and *Text Document – MS-DOS Format* map to `CP_ACP` and `CP_OEM`. But another piece of the puzzle is pretty clear, because MS-DOS used the so-called OEM code page. Therefore, by process of elimination, *Text Document* corresponds to `CP_ACP`. Now that we have puzzled out what the three text formats correspond to, we can address the question "Isn't at least one redundant?" Michael Kaplan explained that ACP and OEM are (usually) different. And neither is the same as Unicode. So in fact all three are (usually) different.

In the United States, the so-called ANSI code page is code page 1252, the so-called OEM code page is code page 437, and Unicode is code page 1200. Here's the string `résumé` expressed in each of the three encodings.

Description	Encoding	Code page (en-us)	Bytes
Text Document	CP_ACP	1252	`72 E9 73 75 6D E9`

Text Document – MS-DOS Format	CP_OEM	437	`72 82 73 75 6D 82`
Unicode Text Document	UTF-16LE	1200	`FF FE 72 00 E9 00 73 00 75 00 6D 00 E9 00`

Three encodings, three different files. No redundancy.

Raymond Chen

Follow