WG 14, n3064

Change request: Writing to multibyte character files

David Svoboda
svoboda@cert.org
Date: 2022-11-10

Here is one undefined behavior entry from Annex J.2. It is #149 in the list of undefined behaviors in J.2:

Use is made of any portion of a file beyond the most recent wide character written to a wide-oriented stream (7.21.2).

This entry originates from C23 (n3047, s7.23.2 p5, 2nd bullet point), which says:

— For wide-oriented streams, after a successful call to a file-positioning function that leaves the file position indicator prior to the end-of-file, a wide character output function can overwrite a partial multibyte character; any file contents beyond the byte(s) written may henceforth not consist of valid multibyte characters.

So first, undefined behavior #149 is wrong. There is no undefined behavior, just the chance of producing a file of invalid multibyte characters according to the current encoding. There also appears to be no relevant text in s7.23.2 (Streams).

Second, the relevant text from p5 is also...well, not wrong, but rather unhelpful. Certainly, if you have an open file for writing, and the file position indicator is not at the end (e.g. it points to some readable bytes), and you write any data, as long at the size of your data leaves some original bytes unchanged in the file, you can convert some valid multibyte characters into invalid multibyte characters, depending on your encoding. That said, it would not be sensible to write wide characters to a file that contains narrow characters, whether multibyte or not.

Next, you don't need to explicitly call a file-positioning function to encounter this problem. You can simply open a pre-existing file for reading and writing (using “r+” as the mode argument to fopen()). The file position indicator starts at the beginning of the file, and you can write anything to it and potentially corrupt pre-existing multibyte characters in the file.

RECOMMENDATION 1: Eliminate undefined behavior #149 from Annex J.2.

RECOMMENDATION 2a: Eliminate 2nd bullet point from s7.23.2 p5.

RECOMMENDATION 2b: Replace 2nd bullet point from s7.23.2 p5 with the following:

- For any input and output stream with pre-existing multibyte characters beyond the file position indicator, any call to write() runs the risk of overwriting part of a multibyte character, leaving the file with bytes that no longer consist of valid multibyte characters.