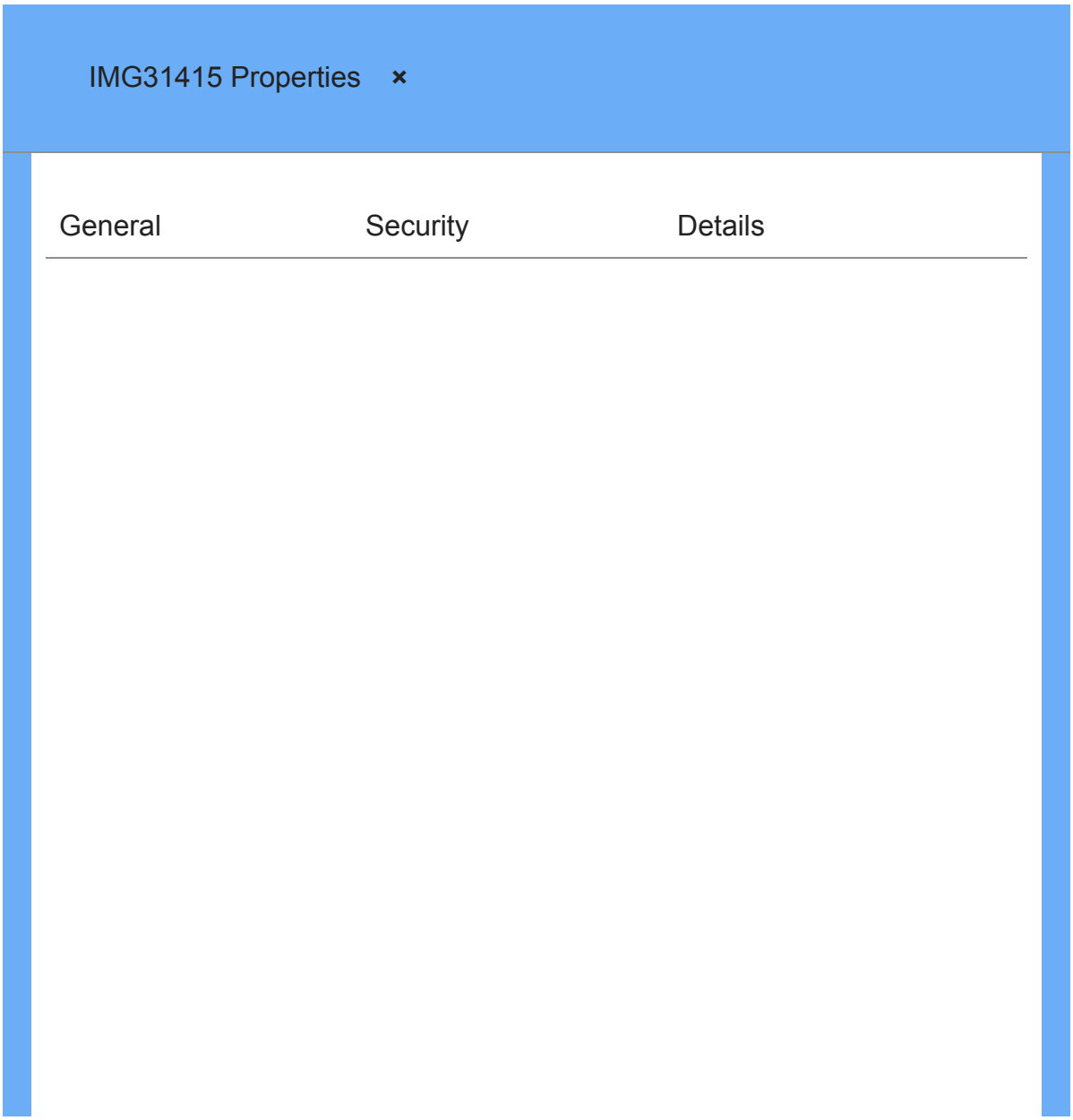
Why is there an invisible U+202A at the start of my file name?

devblogs.microsoft.com/oldnewthing/20150506-00

May 6, 2015

Raymond Chen

There's something strange about this property sheet page:



Okay, that was a trick question, because the thing that's strange is not visible to the eye.

Use the mouse to click in the object name field (the thing with the file path), then press `Home`, followed by `Shift`+`End` to select the entire text, then `Ctrl`+`C` to copy it to the clipboard.

Now things get interesting.

Fire up Notepad, paste the path into the Notepad document, and save it to the desktop with the name `tricky.txt`.

Huh? Notepad says, “This file contains characters in Unicode format which will be lost if you save this as an ANSI encoded text file.”

What Unicode characters are we talking about? There are no accented letters here. All the characters in the file name fit in the ASCII repertoire.

Go to a command prompt and type

```
C:\Users\Bob> copy "
```

and then paste the path from the clipboard, then close the quotation mark, and hit Enter.

```
C:\Users\Bob> copy "?C:\Users\Bob\Desktop\IMG31415.jpg"
The filename, directory name, or volume label syntax is incorrect.
```

Wait, what? Where did that rogue question mark come from?

The answers to the two questions are the same: The mysterious Unicode character, which is invisible in Notepad, and which appears as a question mark on the command line, is U+202A (LEFT-TO-RIGHT EMBEDDING).

We saw some time ago that you can, as a last resort, insert the character U+202B (RIGHT-TO-LEFT EMBEDDING) to force text to be interpreted as right-to-left. The converse character is U+202A (LEFT-TO-RIGHT EMBEDDING), which forces text to be interpreted as left-to-right.

The Security dialog box inserts that control character in the file name field in order to ensure that the path components are interpreted in the expected manner. Unfortunately, it also means that if you try to copy the text out of the dialog box, the Unicode formatting control character comes along for a ride. Since the character is normally invisible, it can create all sorts of silent confusion.

(We’re lucky that the confusion was quickly detected by Notepad and the command prompt. But imagine if you had pasted the path into the source code to a C program!)

Raymond Chen

Follow