Ruby master - Bug #9016

String#encoding is lying?

10/12/2013 04:02 PM - renatosilva (Renato Silva)

Status:	Feedback		
Priority:	Normal		
Assignee:	cruby-windows		
Target version:			
ruby -v:	ruby 2.0.0p247 (2013-06-27) [i386-mingw32]	**Backport:**	1.9.3: UNKNOWN, 2.0.0: UNKNOWN

Description

Please see attached test case.

If you try opening a file using a CP850 (possibly others) path which was passed as command line argument, you are not able at all, unless you encode the argument into its very own reported encoding (CP850), and from some encoding different than that (in my case, both ISO-8859-1 and Windows-1252 worked). It is just like ARGV[0].encoding is lying!

Before, in Ruby 1.8, File.open would work just fine. I have a script that just stopped working, till I found the above workaround. This seems to me like a bug. I would expect Ruby to just do its best in order to convert user input into the required encodings for file APIs and such. Meaning I would not like for a possible fix to require any code migration from 1.8 to 1.9+ at all.

History

#1 - 10/12/2013 08:27 PM - nobu (Nobuyoshi Nakada)

- Status changed from Open to Feedback

- Assignee set to cruby-windows

I know nothing about CP850, give a concrete example path name to reproduce it.

#2 - 10/13/2013 04:29 AM - renatosilva (Renato Silva)

If you type "chcp 850" in cmd.exe before calling the script, it should accept the argument. You can use the word "Japonês" (Japanese) as example for the file path.

#3 - 10/14/2013 06:57 PM - renatosilva (Renato Silva)

- File encoding-lying-reduced.rb added

This reduced test case shows that the argument looks like an ISO-8859-1 string even though its encoding is reported as CP850.

#4 - 10/15/2013 06:34 AM - nobu (Nobuyoshi Nakada)

It would vary on system code pages.

What do you expect and what did you get?

#5 - 10/15/2013 08:35 AM - renatosilva (Renato Silva)

I would expect that if ARGV[0].encoding is CP850, then the string is encoded as CP850. Instead, the string is encoded in another encoding, ISO-8859-1. The reduced test case should output this:

Encoding of argument is reported as CP850 and as valid.
Let us inspect the a-tilde argument: "\xE3"
Let us inspect the a-tilde from UTF-8 source code transcoded into CP850: "\xC6"
Let us inspect the a-tilde from UTF-8 source code transcoded into ISO-8859-1: "\xE3"
RESULT: as you can see, the argument looks like an ISO-8859-1 string, but reports its encoding as CP850.

Files

encoding-lying.rb	1.04 KB	10/12/2013	renatosilva (Renato Silva)
encoding-lying-reduced.rb	761 Bytes	10/14/2013	renatosilva (Renato Silva)