Ruby master - Feature #13124

Should #puts convert to external encoding?

01/11/2017 01:48 PM - Eregon (Benoit Daloze)

Status:	Open	
Priority:	Normal	
Assignee:		
Target version:		

Description

For instance:

```
puts "?\x00\x42\x30".force_encoding(Encoding::UTF_16LE)
?B0
```

```
puts "?\x00\x42\x30".force_encoding(Encoding::UTF_16LE).encode("utf-8")
?▯
```

The first result is surprising to me. It seems to treat the String as raw bytes and just "assume" they are displayable in the external encoding.

Should #puts re-encode the String to print in Encoding.default_external or the locale encoding?

```
STDOUT.set_encoding(Encoding.find("locale"))
```

seems to do what I expect, but should that be the default?

History

#1 - 01/19/2017 06:56 AM - naruse (Yui NARUSE)

On current Ruby, IO converts given string only if the IO object is set internal_encoding.
Therefore the behavior is spec.

Yes, the spec is not clear.
I continually inspecting the use cases and implementation to re-design IO encodings, but it still needs further inspection...

I partially wrote that at https://bugs.ruby-lang.org/issues/7201#note-5

#2 - 01/24/2017 08:55 PM - Eregon (Benoit Daloze)

Thank you for the reply and pointer.

What do you think of having STDOUT, STDERR and STDIN internal_encoding be set by default?
It seems reasonable for those to use the locale encoding.
On the other hand, it seems useless to dump a wide-char String as raw bytes,
it can only be misinterpreted on such a stream.
(Or even more confusing like above where the input is barely related to the actual characters)

Maybe it would be worth to make that an experiment and see what is the impact on compatibility?

#3 - 12/01/2017 04:58 PM - naruse (Yui NARUSE)

- Target version deleted (2.5)

> What do you think of having STDOUT, STDERR and STDIN internal_encoding be set by default?
> It seems reasonable for those to use the locale encoding.
> On the other hand, it seems useless to dump a wide-char String as raw bytes,
> it can only be misinterpreted on such a stream.
> (Or even more confusing like above where the input is barely related to the actual characters)

For STDOUT, it's worth considering.
But it breaks command line utilities which handles encoding other than locale.
I think people who want automatic conversion on STDOUT is not so many to break the compatibility.

#4 - 12/01/2017 06:45 PM - Eregon (Benoit Daloze)

naruse (Yui NARUSE) wrote:

> But it breaks command line utilities which handles encoding other than locale.

What kind of utilities use a different encoding than the locale?
Do you have an example?
It seems to me that if an utility needs to output in a different encoding than the locale,
then one needs to manually set the encoding of STDOUT anyway, so the default encoding conversion would not be an issue.