Ruby - Bug #15718

YAML raises error when dumping strings with UTF32 encoding

03/20/2019 08:21 PM - marcandre (Marc-Andre Lafortune)

Status:	Closed		
Priority:	Normal		
Assignee:	tenderlovemaking (Aaron Patterson)		
Target version:			
ruby -v:	2.6.2p247	**Backport:**	2.4: UNKNOWN, 2.5: UNKNOWN, 2.6: UNKNOWN

Description

```
ruby -r yaml -e "p YAML.dump( ''.force_encoding('UTF-32LE') )"
```

```
Traceback (most recent call last):
 4: from -e:1:in `<main>'
 3: from /Users/work/.rvm/rubies/ruby-2.6.1/lib/ruby/2.6.0/psych.rb:513:in `dump'
 2: from /Users/work/.rvm/rubies/ruby-2.6.1/lib/ruby/2.6.0/psych/visitors/yaml_tree.rb:118:in `
push'
 1: from /Users/work/.rvm/rubies/ruby-2.6.1/lib/ruby/2.6.0/psych/visitors/yaml_tree.rb:136:in `acc
ept'
/Users/work/.rvm/rubies/ruby-2.6.1/lib/ruby/2.6.0/psych/visitors/yaml_tree.rb:298:in `visit_String
': incompatible encoding regexp match (US-ASCII regexp with UTF-32LE string) (Encoding::Compatibil
ityError)
```

Surprisingly, this works in Ruby 2.4.x, but not in 2.2, 2.3, 2.5 nor 2.6!

History

#1 - 03/21/2019 02:26 AM - nobu (Nobuyoshi Nakada)

It may be related to a code range bug.
By adding o.valid_encoding? to Psych::Visitors::YAMLTree#visit_String, the error raises in ruby 24 too.

#2 - 03/21/2019 02:12 PM - rubenochiavone (Ruben Chiavone)

- File yamldumputf32encodingerror.patch added

Since it relates to mismatch of regex and YAML text encoding a possible fix is to only attempt to match the text when encoding matches or when text encoding is ascii_compatible?. WDYT?

Still I'm not sure why on other versions it works.

Anyhow, I'm adding a patch that reproduces and fixes this issues (hopefully).

#3 - 03/21/2019 10:04 PM - marcandre (Marc-Andre Lafortune)

rubenochiavone (Ruben Chiavone) wrote:

> Since it relates to mismatch of regex and YAML text encoding a possible fix is to only attempt to match the text when encoding matches or when text encoding is ascii_compatible?. WDYT?

What about:

```
YAML.dump("Hello\nWorld".encode('UTF-32LE'))
```

or other strings like "123" that need special formatting?

#4 - 03/21/2019 11:33 PM - rubenochiavone (Ruben Chiavone)

I see. There are other regexp based code similar to what Psych::Visitors::YAMLTree.visit_String does. Not sure if testing for encoding before matching as I initially proposed is the way to go. What else do you suggest that could be a fix? Maybe convert it to US_ASCII or skip non-US_ASCII text altogether?

#5 - 03/10/2021 07:11 PM - jeremyevans0 (Jeremy Evans)

- Status changed from Open to Feedback

- Assignee set to tenderlovemaking (Aaron Patterson)

I looked into this and ruby YAML uses libyaml, which is a YAML 1.1 implementation. YAML 1.1 does not support UTF-32 encoding; that isn't supported by the YAML spec until YAML 1.2. So I think it is reasonable for YAML.dump to raise Encoding::CompatibilityError for UTF-32 data, and I don't consider this a bug. Assigning to @tenderlovemaking (Aaron Patterson) to make a decision on whether YAML.dump should handle this.

YAML 1.2 is not backwards compatible with YAML 1.1, so I don't think it would be reasonable to switch the YAML library from libyaml to a different library that supports YAML 1.2. I'm not aware of an existing Ruby library that implements YAML 1.2.

#6 - 03/10/2021 08:04 PM - marcandre (Marc-Andre Lafortune)

Thanks for the investigation @jeremyevans0 (Jeremy Evans) □□

I'm definitely ok to close this. I don't even recall how I stumbled upon this □□

#7 - 03/10/2021 08:07 PM - jeremyevans0 (Jeremy Evans)

- Status changed from Feedback to Closed

Files

yamldumputf32encodingerror.patch	2.55 KB	03/21/2019	rubenochiavone (Ruben Chiavone)