

Ruby - Bug #10598

Cannot make two symbols with same bytes and different encodings

12/15/2014 04:56 AM - DavidEGrayson (David Grayson)

Status:	Closed	Backport: 2.0.0: DONTNEED, 2.1: REQUIRED
Priority:	Normal	
Assignee:		
Target version:		
ruby -v:	ruby 2.2.0preview2 (2014-11-28 trunk 48628) [x86_64-linux]	
Description		
It looks like Ruby 2.1.1 introduced a bug where it is impossible create two different symbols with the same bytes but different encodings. Here is a simple script that reproduces the bug: <pre>sym1 = "ab".force_encoding("UTF-16").to_sym sym2 = "ab".to_sym puts sym2.encoding sym3 = "cd".to_sym sym4 = "cd".force_encoding("UTF-16").to_sym puts sym4.encoding</pre> I would expect the output of this script to be: <pre>US-ASCII UTF-16</pre> The script behaves as expected in Ruby 2.1.0, but in Ruby 2.1.1 and every later version that I tested, it gives incorrect results. Here is a shell session showing the output of the script when I run it in Ruby 2.1.0, 2.1.1, and 2.2.0-preview2: <pre>\$ chruby 2.1.0 && ruby -v && ruby symbol_encoding_bug.rb ruby 2.1.0p0 (2013-12-25 revision 44422) [x86_64-linux] US-ASCII UTF-16 \$ chruby 2.1.1 && ruby -v && ruby symbol_encoding_bug.rb ruby 2.1.1p76 (2014-02-24 revision 45161) [x86_64-linux] UTF-16 US-ASCII \$ chruby 2.2.0-preview2 && ruby -v && ruby symbol_encoding_bug.rb ruby 2.2.0preview2 (2014-11-28 trunk 48628) [x86_64-linux] UTF-16 US-ASCII</pre> It looks like String#to_sym is not properly accounting for the encoding of the string when it searches the symbol table. This is definitely a bug; the value of "ab".to_sym.encoding should be predictable; it should not depend on the state of the symbol table. By the way, JRuby has a similar bug: https://github.com/jruby/jruby/issues/1348		

Associated revisions

Revision 020fcc95fe948f65b03609d445a6cf1bc0c06dc5 - 12/15/2014 07:19 AM - nobu (Nobuyoshi Nakada)

string.c: fix coderange for non-endianness string

- string.c (rb_enc_str_coderange): dummy wchar, non-endianness encoding string cannot be ascii only.
[ruby-core:66835] [Bug #10598]

git-svn-id: svn+ssh://ci.ruby-lang.org/ruby/trunk@48845 b2dd03c8-39d4-4d8f-98ff-823fe69b080e

Revision 020fcc95 - 12/15/2014 07:19 AM - nobu (Nobuyoshi Nakada)

string.c: fix coderange for non-endianness string

- string.c (rb_enc_str_coderange): dummy wchar, non-endianness encoding string cannot be ascii only.
[ruby-core:66835] [Bug #10598]

git-svn-id: svn+ssh://ci.ruby-lang.org/ruby/trunk@48845 b2dd03c8-39d4-4d8f-98ff-823fe69b080e

History

#1 - 12/15/2014 07:19 AM - nobu (Nobuyoshi Nakada)

- Status changed from Open to Closed
- % Done changed from 0 to 100

Applied in changeset r48845.

string.c: fix coderange for non-endianness string

- string.c (rb_enc_str_coderange): dummy wchar, non-endianness encoding string cannot be ascii only.
[\[ruby-core:66835\]](#) [Bug [#10598](#)]

#2 - 01/20/2015 02:42 AM - usa (Usaku NAKAMURA)

- Backport changed from 2.0.0: UNKNOWN, 2.1: UNKNOWN to 2.0.0: DONTNEED, 2.1: REQUIRED