Ruby - Bug #5486

rb_stat() doesn't respect input encoding

10/26/2011 10:00 PM - now (Nikolai Weibull)

Status:	Closed		
Priority:	Normal		
Assignee:	nobu (Nobuyoshi Nakada)		
Target version:			
ruby -v:	-	**Backport:**	

Description

rb_stat() overrides the input strings encoding and applies one of various encodings through rb_str_encode_ospath(). This may be convenient for certain kinds of user input or input from a source file in a different encoding, but it isn't good for other kinds of user input or input from other functions, such as Dir.entries.

If Ruby wants us to be explicit about encodings, then Ruby shouldn't change it behind our backs.

I suspect that this is an issue that may appear in various other functions as well.

History

#1 - 10/28/2011 02:28 PM - usa (Usaku NAKAMURA)

Sorry, I can't understand your point.
If you think there is a bug, would you show us the bug by code?

#2 - 10/28/2011 03:23 PM - now (Nikolai Weibull)

On Fri, Oct 28, 2011 at 07:28, Usaku NAKAMURA redmine@ruby-lang.org wrote:

> Sorry, I can't understand your point.
> If you think there is a bug, would you show us the bug by code?

That's hard to do, but name a file in an encoding other than
'filesystem' on an NTFS filesystem. What I did was accidentally
create a file whose name was encoded in UTF-16. Then, do
Dir['dir'].entries.each{ |e| printf "%p: %s\n", e, File.file? e },
where 'dir' is the directory containing this file. e.file? will
return false for this file, even though it's a file. The problem is,
as explained, in rb_stat(), as it re-encodes its argument in the
'filesystem' encoding.

#3 - 10/28/2011 03:53 PM - now (Nikolai Weibull)

On Fri, Oct 28, 2011 at 08:14, Nikolai Weibull now@bitwi.se wrote:

> On Fri, Oct 28, 2011 at 07:28, Usaku NAKAMURA redmine@ruby-lang.org wrote:
>
> > Sorry, I can't understand your point.
> > If you think there is a bug, would you show us the bug by code?
>
> That's hard to do, but name a file in an encoding other than
> 'filesystem' on an NTFS filesystem. What I did was accidentally
> create a file whose name was encoded in UTF-16. Then, do
> Dir['dir'].entries.each{ |e| printf "%p: %s\n", e, File.file? e },
> where 'dir' is the directory containing this file. e.file? will
> return false for this file, even though it's a file. The problem is,
> as explained, in rb_stat(), as it re-encodes its argument in the
> 'filesystem' encoding.

Actually, it's probably easier than that. It can be done on a HFS+
filesystem (and probably any other, as well) just as easily

```
% echo $LC_CTYPE
UTF-8
% mkdir t
```

```
% touch t/å
% cat > a.rb
```

-- coding: utf-8 --

```
Dir.new('t').entries.each{ |e| printf "%p, %p, %s\n", e, e.encoding,
File.file?(e) }
^D
% ruby --version
ruby 2.0.0dev (2011-10-26 trunk 33526) [x86_64-darwin10.8.0]
% ruby a.rb
".", #Encoding:UTF-8, false
"..", #Encoding:UTF-8, false
"å", #Encoding:UTF-8, false
```

I guess the problem is that Ruby assumes that it can apply an encoding
to something that it gets from the filesystem when it would probably
be better to not do so. It should probably be BINARY or ASCII-8BIT
instead of UTF-8.

(It turns out that this example gave the same results in 1.8.7 (minus
the e.encoding), so perhaps I'm doing something else wrong.)

Trying to do

p File.file?('t/å'.encode('UTF-16LE'))

results in

in `file?': path name must be ASCII-compatible (UTF-16LE): "t/\u00E5"
(Encoding::CompatibilityError)

I give up.

#4 - 10/28/2011 04:23 PM - nobu (Nobuyoshi Nakada)

- ruby -v changed from ruby 1.9.3dev (2011-09-13 revision 33263) [i386-mingw32] to -

Hi,

(11/10/28 15:35), Nikolai Weibull wrote:

> Actually, it's probably easier than that. It can be done on a HFS+
> filesystem (and probably any other, as well) just as easily

It's not true.

```
    % echo $LC_CTYPE
    UTF-8
    % mkdir t
    % touch t/å
    % cat > a.rb
```

-- coding: utf-8 --

```
    Dir.new('t').entries.each{ |e| printf "%p, %p, %s\n", e, e.encoding,
    File.file?(e) }
    ^D
```

`e' doesn't have directory prefix, "t/". It can't stat.

```
$ ruby -v -C t -e 'Dir.foreach(".") {|e| printf "%p, %p, %p\n", e, e.encoding, File.file?(e)}'
ruby 2.0.0dev (2011-10-25 trunk 33523) [universal.x86_64-darwin11.2.0]
".", #Encoding:UTF-8, false
"..", #Encoding:UTF-8, false
"å", #Encoding:UTF-8, true
```

--
Nobu Nakada

#5 - 10/28/2011 07:23 PM - now (Nikolai Weibull)

On Fri, Oct 28, 2011 at 09:20, Nobuyoshi Nakada nobu@ruby-lang.org wrote:

> (11/10/28 15:35), Nikolai Weibull wrote:
>
>> Actually, it's probably easier than that. It can be done on a HFS+
>> filesystem (and probably any other, as well) just as easily
>
>
> It's not true.
>
>> % echo $LC_CTYPE
>> UTF-8
>> % mkdir t
>> % touch t/å
>> % cat > a.rb
>
> # *-- coding: utf-8 –*
>
>> Dir.new('t').entries.each{ |e| printf "%p, %p, %s\n", e, e.encoding,
>> File.file?(e) }
>> ^D
>
>
> `e' doesn't have directory prefix, "t/". It can't stat.

Ouch, of course. How stupid of me. That explains why it didn't work
under 1.8.7 either.

The point still remains valid on Windows, however:

% mkdir t
% touch t/→
% ruby -v -C t -e 'Dir.foreach(".") {|e| printf "%p, %p, %p\n", e,
e.encoding, File.file?(e)}'
ruby 1.9.3dev (2011-09-13 revision 33263) [i386-mingw32]
".", #Encoding:Windows-1252, false
"..", #Encoding:Windows-1252, false
"?", #Encoding:Windows-1252, false

Hm, I guess here the result of Dir.foreach is broken.

Here's another case:

% ruby -v -rfind -e 'Find.find("t").each{ |e| printf "%p, %s, %p,
%p\n", e, e.dump, e.encoding, File.file?(e)}'
"t", "t", #Encoding:UTF-8, false
"t/?", "t/?", #Encoding:ASCII-8BIT, false

Equally broken, I guess.

#6 - 03/11/2012 04:27 PM - ko1 (Koichi Sasada)

- Status changed from Open to Assigned

- Assignee set to nobu (Nobuyoshi Nakada)

#7 - 03/12/2012 06:41 AM - nobu (Nobuyoshi Nakada)

- Category changed from core to M17N

- Status changed from Assigned to Feedback

- Priority changed from 5 to 3

Does this issue still occur?

#8 - 03/13/2012 06:23 PM - now (Nikolai Weibull)

On Sun, Mar 11, 2012 at 22:41, Nobuyoshi Nakada nobu@ruby-lang.org wrote:

> Issue #5486 has been updated by Nobuyoshi Nakada.

Category changed from core to M17N
Status changed from Assigned to Feedback
Priority changed from High to Low

Does this issue still occur?

Yes, it still occurs against trunk:

ruby 1.9.3dev (2011-09-13 revision 33263) [i386-mingw32]

#9 - 03/15/2012 01:29 PM - now (Nikolai Weibull)

2012/3/15 U.Nakamura usa@garbagecollect.jp:

> Hello,
>
> In message "[ruby-core:43260] Re: [ruby-core:43236] [ruby-trunk - Bug #5486][Feedback] rb_stat() doesn't respect input encoding"
>   on Mar.13,2012 18:03:04, now@bitwi.se wrote:
>
>> Yes, it still occurs against trunk:
>>
>> ruby 1.9.3dev (2011-09-13 revision 33263) [i386-mingw32]
>
> It's not trunk...
> It seems too old.

How can you say that? I just tested it and got the same results. I
showed you my version string above, that's trunk.

#10 - 03/15/2012 02:23 PM - Anonymous

On 3/14/12 11:24 PM, Nikolai Weibull wrote:

> 2012/3/15 U.Nakamurausa@garbagecollect.jp:
>
>> Hello,
>>
>> In message "[ruby-core:43260] Re: [ruby-core:43236] [ruby-trunk - Bug #5486][Feedback] rb_stat() doesn't respect input encoding"
>>   on Mar.13,2012 18:03:04,now@bitwi.se wrote:
>>
>>> Yes, it still occurs against trunk:
>>>
>>> ruby 1.9.3dev (2011-09-13 revision 33263) [i386-mingw32]
>>> It's not trunk...
>>> It seems too old.
>>> How can you say that? I just tested it and got the same results. I
>>> showed you my version string above, that's trunk.

hi Nikolai,

try compiling an updated version of trunk from this repository:
svn co http://svn.ruby-lang.org/repos/ruby/branches/ruby_1_9_3

your version indicates it's from last year. here's a version string from
a recent compilation on my system:
% ./ruby --version
ruby 1.9.3p163 (2012-03-14 revision 35012) [x86_64-darwin11.3.0]

does that help at all?

#11 - 03/15/2012 02:59 PM - naruse (Yui NARUSE)

2012/3/15 Trevor Wennblom trevor@well.com:

> try compiling an updated version of trunk from this repository:
> svn co http://svn.ruby-lang.org/repos/ruby/branches/ruby_1_9_3

It is ruby_1_9_3 branch, not trunk.

For trunk,
svn co http://svn.ruby-lang.org/repos/ruby/trunk

--
NARUSE, Yui naruse@airemix.jp

#12 - 03/15/2012 04:43 PM - naruse (Yui NARUSE)

- Status changed from Feedback to Closed

#13 - 03/15/2012 04:53 PM - now (Nikolai Weibull)

On Thu, Mar 15, 2012 at 05:24, Nikolai Weibull now@bitwi.se wrote:

> 2012/3/15 U.Nakamura usa@garbagecollect.jp:
>
>> Hello,
>>
>> In message "[ruby-core:43260] Re: [ruby-core:43236] [ruby-trunk - Bug #5486][Feedback] rb_stat() doesn't respect input encoding" on Mar.13,2012 18:03:04, now@bitwi.se wrote:
>>
>>> Yes, it still occurs against trunk:
>>>
>>> ruby 1.9.3dev (2011-09-13 revision 33263) [i386-mingw32]
>>
>>
>> It's not trunk...
>> It seems too old.
>
>
> How can you say that? I just tested it and got the same results. I
> showed you my version string above, that's trunk.

Argh, sorry. I ran the test with the incorrect PATH, after all. Yes,
this issue has been resolved. You can close it.