Ruby master - Bug #4010
YAML fails to roundtrip non ASCII String

11/01/2010 11:36 AM - phasis68 (Heesob Park)

<table>
<thead>
<tr>
<th>Status:</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority:</td>
<td>Normal</td>
</tr>
<tr>
<td>Assignee:</td>
<td></td>
</tr>
<tr>
<td>Target version:</td>
<td>2.0.0</td>
</tr>
<tr>
<td>ruby -v:</td>
<td>ruby 1.9.3dev (2010-11-01 trunk 29655) [i386-mswin32_90]</td>
</tr>
<tr>
<td>Backport:</td>
<td></td>
</tr>
</tbody>
</table>

Description
=begin
C:>ruby -v -ryaml -e 's="한글";p YAML.load(YAML.dump(s))==s'
ruby 1.9.3dev (2010-11-01 trunk 29655) [i386-mswin32_90]
false
C:>ruby -v -ryaml -e 's="한글";p YAML.load(YAML.dump(s))==s'
ruby 1.8.6 (2010-02-04 patchlevel 398) [i386-mingw32]
true
=end

History
#1 - 11/02/2010 08:54 PM - tenderlovemaking (Aaron Patterson)
=begin
On Mon, Nov 01, 2010 at 11:36:58AM +0900, Heesob Park wrote:

Bug #4010: YAML fails to roundtrip non ASCII String
http://redmine.ruby-lang.org/issues/show/4010

Author: Heesob Park
Status: Open, Priority: Normal
Category: lib, Target version: 1.9.x
ruby -v: ruby 1.9.3dev (2010-11-01 trunk 29655) [i386-mswin32_90]

C:>ruby -v -ryaml -e 's="한글";p YAML.load(YAML.dump(s))==s'
ruby 1.9.3dev (2010-11-01 trunk 29655) [i386-mswin32_90]
false
C:>ruby -v -ryaml -e 's="한글";p YAML.load(YAML.dump(s))==s'
ruby 1.8.6 (2010-02-04 patchlevel 398) [i386-mingw32]
true

I'm pretty sure this is a known issue with Syck. Can you try to reproduce this with psych?

ruby -v -ryaml -e "YAML::ENGINE.yamler = "psych"; s="한글";p YAML.load(YAML.dump(s))==s"

--
Aaron Patterson
http://tenderlovemaking.com/

Attachment: (unnamed)
=end

#2 - 11/02/2010 09:58 PM - phasis68 (Heesob Park)
=begin
The same result with psych.

$ ruby -v -ryaml -e "YAML::ENGINE.yamler = "psych"; s="한글";p YAML.load(YAML.dump(s))==s"
ruby 1.9.3dev (2010-11-02 trunk 29667) [i686-linux]
/usr/local/lib/ruby/1.9.1/psych/deprecated.rb:79: warning: method redefined; discarding old to_yaml_properties
/usr/local/lib/ruby/1.9.1/syck/rubytypes.rb:13: warning: previous definition of to_yaml_properties was here
false

09/20/2021
FYI the current encoding is 'EUC-KR'.
I know it works when the encoding is 'UTF-8'.

=end

#3 - 11/03/2010 11:21 AM - tenderlovemaking (Aaron Patterson)

=begin
On Tue, Nov 02, 2010 at 09:58:27PM +0900, Heesob Park wrote:

Issue #4010 has been updated by Heesob Park.

The same result with psych.

$ ruby -v -ryaml -e 'YAML::ENGINE.yamler = "psych"; s="한글";pYAML.load(YAML.dump(s)) == s'
ruby 1.9.3dev (2010-11-02 trunk 29667) [i686-linux]
/usr/local/lib/ruby/1.9.1/psych/deprecated.rb:79: warning: method redefined; discarding old to_yaml_properties
/usr/local/lib/ruby/1.9.1/syck/rubytypes.rb:13: warning: previous definition of to_yaml_properties was here
false
FYI the current encoding is 'EUC-KR'.
I know it works when the encoding is 'UTF-8'.

I think the problem is that your default_internal encoding isn't being set. YAML is usually stored as UTF-8, but psych will automatically transcode the string to whatever your default_internal encoding is set to.

Maybe this script will help illustrate the problem:

```ruby
# coding: utf-8
require 'psych'
eucjp = "한글".encode('EUC-JP')
string = Psych.load(Psych.dump(eucjp))

p string.encoding #=> #<Encoding:UTF-8>
p eucjp == string #=> false

Encoding.default_internal = 'EUC-JP'

string = Psych.load(Psych.dump(eucjp))
p string.encoding #=> #<Encoding:EUC-JP>
p eucjp == string #=> true
```

Try running your Ruby like this:

$ ruby -EEUC-KR:EUC-KR -v -ryaml -e 'YAML::ENGINE.yamler = "psych"; s="한글";pYAML.load(YAML.dump(s)) == s'

Aaron Patterson
http://tenderlovemaking.com/

 Attachment: (unnamed)
=end

#4 - 11/03/2010 11:36 AM - phasis68 (Heesob Park)

=begin
2010/11/3 Aaron Patterson
aaron@tenderlovemaking.com:

On Tue, Nov 02, 2010 at 09:58:27PM +0900, Heesob Park wrote:

Issue #4010 has been updated by Heesob Park.

The same result with psych.

$ ruby -v -ryaml -e 'YAML::ENGINE.yamler = "psych"; s="한글";pYAML.load(YAML.dump(s)) == s'
ruby 1.9.3-dev (2010-11-02 trunk 29667) [i686-linux]
/usr/local/lib/ruby/1.9.1/psych/deprecated.rb:79: warning: method redefined; discarding old to_yaml_properties
/usr/local/lib/ruby/1.9.1/syck/rubytypes.rb:13: warning: previous definition of to_yaml_properties was here
false
FYI the current encoding is 'EUC-KR'.

09/20/2021
I think the problem is that your default_internal encoding isn't being set. YAML is usually stored as UTF-8, but psych will automatically transcode the string to whatever your default_internal encoding is set to.

Maybe this script will help illustrate the problem:

```ruby
# coding: utf-8
require 'psych'
eucjp = "こんにちは!".encode('EUC-JP')
string = Psych.load(Psych.dump(eucjp))
p stringencoding #=>_ENCODING:UTF-8
p eucjp == string #=> false

Encoding.default_internal = 'EUC-JP'
string = Psych.load(Psych.dump(eucjp))
p stringencoding #=>_ENCODING:EUC-JP
p eucjp == string #=> true

Try running your Ruby like this:

$ ruby -EEUC-KR:EUC-KR -v -ryaml -e 'YAML::ENGINE.yamler = "psych"; s="한글";pYAML.load(YAML.dump(s))===s'

Here is the result

$ ruby -EEUC-KR:EUC-KR -v -ryaml -e 'YAML::ENGINE.yamler = "psych"; s="한글";pYAML.load(YAML.dump(s))===s'
ruby 1.9.3dev (2010-11-02 trunk 29667) [i686-linux]
true

Did you mean it is not a bug and I must specify the default external and internal character encodings?

$ irb
irb(main):001:0> Encoding.default_external
=> #ENCODING:EUC-KR
irb(main):002:0> Encoding.default_internal
=> nil

Why ruby cannot detect Encoding.default_internal ?

Regards,
Park Heesob

=end

#5 - 11/04/2010 01:02 AM - tenderlovemaking (Aaron Patterson)
=begin
On Wed, Nov 03, 2010 at 11:36:35AM +0900, Heesob Park wrote:
2010/11/3 Aaron Patterson aaron@tenderlovemaking.com:

On Tue, Nov 02, 2010 at 09:58:27PM +0900, Heesob Park wrote:

Issue #4010 has been updated by Heesob Park.
The same result with psych.

$ ruby -v -ryaml -e "YAML::ENGINE.yamler = "psych"; s="한글";pYAML.load(YAML.dump(s))===s"
ruby 1.9.3dev (2010-11-02 trunk 29667) [i686-linux]
/vendor/lib/ruby/1.9.1/psych/deprecated.rb:79: warning: method redefined; discarding old to_yaml_properties
/vendor/lib/ruby/1.9.1/syck/rubytypes.rb:13: warning: previous definition of to_yaml_properties was here
false

FYI the current encoding is 'EUC-KR'.
I know it works when the encoding is 'UTF-8'.
```
I think the problem is that your default_internal encoding isn’t being set. YAML is usually stored as UTF-8, but psych will automatically transcode the string to whatever your default_internal encoding is set to.

Maybe this script will help illustrate the problem:

```ruby
# coding: utf-8
require 'psych'
eucjp = “こんにちは!”.encode('EUC-JP')
string = Psych.load(Psych.dump(eucjp))

p string.encoding # => Encoding:UTF-8
p eucjp == string # => false

Encoding.default_internal = ‘EUC-JP’

string = Psych.load(Psych.dump(eucjp))

p string.encoding # => Encoding:EUC-JP
p eucjp == string # => true
```

Try running your Ruby like this:

```
$ ruby -EEUC-KR:EUC-KR -v -ryaml -e 'YAML::ENGINE.yamler = “psych”; s=“한글”;pYAML.load(YAML.dump(s))===s'
```

Here is the result:

```
true
```

Did you mean it is not a bug and I must specify the default external and internal character encodings?

Yes. YAML is stored as UTF-8 or 16 (possibly 32 as well, depending on the spec you choose). Strings you pull out of Psych will have an encoding that matches the source document.

```
$ irb
iri(main):001:0> Encoding.default_external
=> Encoding:EUC-KR
iri(main):002:0> Encoding.default_internal
=> nil
```

"default_external" indicates the encoding that files on disk probably have. A good default is the OS setting.

"default_internal" indicates the encoding that you want strings internal to your programs to have. Making that decision is not so easy. Magic comments cannot be used because there can exist many magic comments in your programs.

"default_internal" should be used by things like database adapters (or in this case YAML parsers) where the encoding of the external entity may differ from the encoding the user wants to use. The external entity should transcode to the user's "default_internal" setting. Because of this logic, there exists another problem with setting "default_internal" to something for the user.

Here is an example of what could happen if default_internal was automatically set:

"default_internal" is automatically set to something, say Shift-JIS, and you load a YAML file. You know that your YAML file is stored as UTF-8, and yet when you output the encoding from your program, you're surprised to see it report Shift-JIS! You wanted the original encoding of the file, but now you must call encode() to get it back to UTF-8. Even worse,
because you went from UTF-8 => Shift-JIS => UTF-8, now there may be data loss due to encoding round trip problems.

Hope that helps!

--
Aaron Patterson
http://tenderlovemaking.com/
Attachment: (unnamed)
=end

#6 - 11/05/2010 12:16 AM - tenderlovemaking (Aaron Patterson)
- Status changed from Open to Rejected

=begin
Closing since Psych works as expected. "default_internal" must be set for strings to be automatically transcoded.
=end