

Ruby - Feature #12650

Use UTF-8 encoding for ENV on Windows

08/03/2016 12:53 AM - davispuh (Dāvis Mosāns)

Status:	Closed	
Priority:	Normal	
Assignee:	windows	
Target version:	3.0	
Description Windows environment variables supports Unicode (same wide WinAPI) and so there's no reason to limit ourselves to any codepage. Currently ENV would use locale's encoding (console's codepage) which obviously won't work correctly for characters outside of those codepages. I've attached a patch which implements this and fixes bug #9715		
Related issues: Related to Ruby - Bug #9715: ENV data yield ASCII-8BIT encoded strings under ... Closed Related to Ruby - Feature #16604: Set default for Encoding.default_external t... Closed		

History

#1 - 08/03/2016 05:36 AM - usa (Usaku NAKAMURA)

We don't want to break compatibility.
Wait Ruby3.

#2 - 08/03/2016 06:06 AM - nobu (Nobuyoshi Nakada)

- Tracker changed from Bug to Feature

#3 - 08/03/2016 07:46 PM - spatulasnout (B Kelly)

Hi,

Usaku NAKAMURA wrote:

We don't want to break compatibility.
Wait Ruby3.

We always invoke ruby with -EUTF-8:UTF-8 .

Would make sense to enable this patch in ruby 2.x in such situations where UTF-8 behavior has been requested explicitly?

#4 - 09/12/2016 06:38 AM - naruse (Yui NARUSE)

- Related to Bug #9715: ENV data yield ASCII-8BIT encoded strings under Windows with unicode username added

#5 - 10/08/2016 01:25 AM - Iristyle (Ethan Brown)

If you could rethink the plan to wait until Ruby 3, that would be great.

I would expect Ruby to normalize on UTF-8 strings everywhere internally, and only convert to local codepage on the boundary (such as writing to console, file, etc).

We are tracking a number of issues in Puppet that we believe are caused by the current behavior:

- [Puppet Throws Exception when Running Under Unicode Windows User](#)
- [Bundler Fails when Running Under a Unicode Windows User](#)
- [Puppet Crashes when Unicode User Applies Manifest](#)

#6 - 01/04/2017 02:15 PM - thomthom (Thomas Thomassen)

B Kelly wrote:

Hi,

Usaku NAKAMURA wrote:

We don't want to break compatibility.
Wait Ruby3.

We always invoke ruby with `-EUTF-8:UTF-8`.

Would make sense to enable this patch in ruby 2.x in such situations where UTF-8 behavior has been requested explicitly?

I would like to second this request. We are also troubled by the encoding issues under Windows. Not sure when Ruby 3 is planned to be released, but we would prefer for a more immediate solution.

#7 - 03/13/2017 03:02 PM - shyouhei (Shyouhei Urabe)

We looked at this issue in today's developer meeting.

First off, attendees' understanding: ENV in Windows is managed by its kernel, and is provided to an userland process as an array of wide characters. Tell me if it's wrong. Also, we already support writing UTF_8 strings into ENV because that has no backwards compatibility problem. The problem is to read from it.

Now, from our long tradition of using OEM codepage in Windows, it has been difficult to change the encoding of ENV to UTF_8. A tragedy is Windows does have chcp 65001, which is not practically used anywhere. So windows users are left in their code pages.

I understand you want to use UTF_8. In order to do so, changing default encoding is not practically possible now because of backwards compatibility. I advice you to propose other ways; like for instance having some sort of "UTF_8 mode"-like thing. Maybe does it make sense for you to set default_internal encoding (which is set to nil by default)?

#8 - 03/13/2017 04:33 PM - thomthom (Thomas Thomassen)

I would be ok with it not being default, as long as it can be configured for the whole interpreter and not some magic comment that would have to be in each source file.

In our particular scenario we are embedding Ruby into our application and we would like to configure the Ruby interpreter to use this "UTF-8 mode". People that are writing Ruby extensions for our application already have to use hacks such as force_encoding to correct this - and it's a constant source of bugs and problems. If we could force ENV strings to be UTF-8 by default for the embedded environment we provide that be a great relief for us.

shyouhei (Shyouhei Urabe) wrote:

We looked at this issue in today's developer meeting.

First off, attendees' understanding: ENV in Windows is managed by its kernel, and is provided to an userland process as an array of wide characters. Tell me if it's wrong. Also, we already support writing UTF_8 strings into ENV because that has no backwards compatibility problem. The problem is to read from it.

Now, from our long tradition of using OEM codepage in Windows, it has been difficult to change the encoding of ENV to UTF_8. A tragedy is Windows does have chcp 65001, wich is not practically used anywhere. So windows users are left in their code pages.

I understand you want to use UTF_8. In order to do so, changing default encoding is not practically possible now because of backwards compatibility. I advice you to propose other ways; like for instance having some sort of "UTF_8 mode"-like thing. Maybe does it make sense for you to set default_internal encoding (which is set to nil by default)?

#9 - 12/26/2019 05:56 AM - naruse (Yui NARUSE)

- Assignee set to windows

- Target version set to 3.0

#10 - 01/14/2020 09:31 AM - naruse (Yui NARUSE)

- Tags set to win, encoding

#11 - 11/26/2020 12:26 PM - larskanis (Lars Kanis)

A patch for ruby-3.0 is here: <https://github.com/ruby/ruby/pull/3818>

#12 - 12/07/2020 04:49 PM - naruse (Yui NARUSE)

- Related to Feature #16604: Set default for Encoding.default_external to UTF-8 on Windows added

#13 - 12/09/2020 08:10 AM - larskanis (Lars Kanis)

This issue can be closed. It's merged in [ca76337a00244635faa331afd04f4b75161ce6fb](https://github.com/ruby/ruby/commit/ca76337a00244635faa331afd04f4b75161ce6fb)

#14 - 12/09/2020 08:50 AM - duerst (Martin Dürst)

- Status changed from Open to Closed

Files

0001-Always-use-UTF-8-encoded-environment-on-Windows.patch	3.64 KB	08/03/2016	davispuh (Dāvis Mosāns)
--	---------	------------	-------------------------