Ruby master - Feature #13488
Set Encoding.default_external to UTF-8 on Windows

04/20/2017 11:32 AM - larskanis (Lars Kanis)

Status: Closed
Priority: Normal
Assignee:
Target version: 3.0

Description
Currently Encoding.default_external is set to the local ANSI encoding of the Windows installation unless changed per option -E. This is cp850 for Western Europe. It should be changed to UTF-8.

The current setting is a major interoperability issue and it is neither useful nor expected, because nobody seriously uses the ancient locale dependent cpXYZ encodings for file content.

If a native encoding shall be used, it should be UTF-16 on Windows. However UTF-16 would make interoperability and compatibility even more difficult. So the only reliable choice for default_external is UTF-8, IMHO.

This is already patched per [1] in the upcoming RubyInstaller-2.4 for Windows release. It additionally requires a few changes to MRIs encoding tests.


Related issues:
Related to Ruby master - Feature #16604: Set default for Encoding.default_ext...

History
#1 - 04/20/2017 12:40 PM - usa (Usaku NAKAMURA)
larskanis (Lars Kanis) wrote:

The current setting is a major interoperability issue and it is neither useful nor expected, because nobody seriously uses the ancient locale dependent cpXYZ encodings for file content.

Come Japan and check Japanese Windows users' text files :)”

If a native encoding shall be used, it should be UTF-16 on Windows. However UTF-16 would make interoperability and compatibility even more difficult. So the only reliable choice for default_external is UTF-8, IMHO.

I agree.

We won't change the default of Encoding.default_encoding before Ruby3 because of to keep compatibility.
And, at Ruby3, we will change it to UTF-8, perhaps.

#2 - 04/21/2017 06:05 AM - naruse (Yui NARUSE)
- Status changed from Open to Rejected

Though it should be changed into UTF-8 in the future, I don't plan it in 2.5 too.
(If many people want to change, I may change my mind)

Anyway the patch is directly changes the default_external, but default_external's default is correctly locale.
The locale encoding affects not only default_external, it also affects filesystem encoding.

If you want to change them, you should just change ruby as if it runs on chcp 65001:

diff --git a/localeinit.c b/localeinit.c
index fa9cc26b8e..397b12bf6c 100644
--- a/localeinit.c
+++ b/localeinit.c
@@ -126,8 +126,7 @@ Init_enc_set_filesystem_encoding(void)
     idx = ENCINDEX_US_ASCII;
 #elif defined _WIN32
     char cp[SIZEOF_CP_NAME];

09/18/2022
const UINT codepage = ruby_w32_codepage ? ruby_w32_codepage :
    AreFileApisANSI() ? GetACP() : GetOEMCP();
+ const UINT codepage = 65001;
    CP_FORMAT(cp, codepage);
    idx = rb_enc_find_index(cp);
    if (idx < 0) idx = ENCINDEX_ASCII;

#3 - 04/26/2017 01:21 PM - MSP-Greg (Greg L)
usa (Usaku NAKAMURA) wrote:

    Come Japan and check Japanese Windows users’ text files ;)

On the topic of Japanese Windows users, what type of Ruby builds are they using, mswin or MinGW?

#4 - 04/26/2017 01:44 PM - usa (Usaku NAKAMURA)
MSP-Greg (Greg L) wrote:

    On the topic of Japanese Windows users, what type of Ruby builds are they using, mswin or MinGW?

Of course, we use both, as far as I know.

#5 - 04/26/2017 01:56 PM - MSP-Greg (Greg L)
usa (Usaku NAKAMURA) wrote:

    MSP-Greg (Greg L) wrote:

        On the topic of Japanese Windows users, what type of Ruby builds are they using, mswin or MinGW?

        Of course, we use both, as far as I know.

        Sorry for not being more specific. As far as you know, approximately what percentage of Japanese Windows Ruby users use MinGW builds, and what percentage use mswin builds?

#6 - 04/26/2017 02:09 PM - usa (Usaku NAKAMURA)
MSP-Greg (Greg L) wrote:

    Sorry for not being more specific. As far as you know, approximately what percentage of Japanese Windows Ruby users use MinGW builds, and what percentage use mswin builds?

Sorry, I don't know such statistics.

#7 - 05/01/2017 09:51 AM - duerst (Martin Dürst)
larskanis (Lars Kanis) wrote:

    Currently Encoding.default_external is set to the local ANSI encoding of the Windows installation unless changed per option -E. This is cp850 for Western Europe. It should be changed to UTF-8.

    The current setting is a major interoperability issue and it is neither useful nor expected, because nobody seriously uses the ancient locale dependent cpXYZ encodings for file content.

I wouldn't say that it is a setting that it totally without problems (but see below).

    If a native encoding shall be used, it should be UTF-16 on Windows. However UTF-16 would make interoperability and compatibility even more difficult. So the only reliable choice for default_external is UTF-8, IMHO.

I would strongly suggest to backpedal on [1]. As a longtime supporter of UTF-8, I have on various occasions tried to use UTF-8 on a Japanese Windows. I just tried again (on Windows 8.1, your mileage may vary). Unfortunately, it only works partially, not good enough to be worth it. I detected two problems that essentially make it unusable:

1. When using chcp 65001, the IME stops working. This means that it’s no longer possible to type anything except ASCII on the keyboard.

2. Output from commands gets garbled. As an example, more produces garbage (probably Shift_JIS interpreted as UTF-8). The user may be able to live with garbage, but the problem is that the garbage also messes up the command itself; after a few times of typing the Enter key, I get a
small window saying "More Utility has stopped working".

This is already patched per [1] in the upcoming RubyInstaller-2.4 for Windows release. It additionally requires a few changes to MRIs encoding tests.

Please don't make this change, or make it only for those versions and code pages of Windows where you are sure that the negative consequences (see above) won't outweigh the advantages.


As for what I'm using on Windows, it's mostly cygwin. But many of my students have used the Ruby Installer in the past, successfully. It would be a pity if it were no longer possible to use the Ruby Installer on Japanese Windows.

#8 - 05/01/2017 11:22 AM - nobu (Nobuyoshi Nakada)
naruse (Yui NARUSE) wrote:

If you want to change them, you should just change ruby as if it runs on chcp 65001:

In the trunk, you don't have to change ruby source code.
Just configure with --enable-debug-env, and include codepage=65001 in the environment variable RUBY_DEBUG.

C:\Users\nobu\work\ruby\trunk\x64-mswin32_140>.in\ruby -e "p Encoding.default_external, Encoding.find('filesystem')"
#<Encoding:Windows-31J>
#<Encoding:Windows-31J>
C:\Users\nobu\work\ruby\trunk\x64-mswin32_140>set RUBY_DEBUG=codepage=65001
C:\Users\nobu\work\ruby\trunk\x64-mswin32_140>.in\ruby -e "p Encoding.default_external, Encoding.find('filesystem')"
#<Encoding:UTF-8>
#<Encoding:UTF-8>

#9 - 05/09/2017 08:41 PM - Iristyle (Ethan Brown)
I agree that changing the default to UTF-8 is not appropriate on Windows in a Ruby 2.x release.
Should a change occur in Ruby 3 to make UTF-8 the default, I believe it would still be useful to gain access to the original Windows codepage as this may be the encoding of files.
I was similarly surprised to see that the rubyinstaller2 project had made the change to make UTF-8 the default on Windows in the Ruby 2.4 installer builds, and I've filed an issue at https://github.com/oneclick/rubyinstaller2/issues/38 in hopes that it will be reverted (another user has filed a similar issue report at https://github.com/oneclick/rubyinstaller2/issues/37)

#10 - 12/07/2020 04:43 PM - naruse (Yui NARUSE)
- Status changed from Rejected to Open
- Target version set to 3.0

#11 - 12/07/2020 04:43 PM - naruse (Yui NARUSE)
- Related to Feature #16604: Set default for Encoding.default_external to UTF-8 on Windows added

#12 - 12/07/2020 04:44 PM - naruse (Yui NARUSE)
- Status changed from Open to Closed