On Windows use UTF-8 as filesystem encoding

Windows (NTFS) supports Unicode and there can be paths/filenames with other characters than current ANSI/OEM codepage can encode.

See attached patch.

Revision 5b98b2ce - 12/20/2020 09:34 AM - nobu (Nobuyoshi Nakada)
win32: Use UTF-8 as filesystem encoding [Feature #12654]
Co-Authored-By: Dāvis Mosāns davispuh@gmail.com

Revision 95862ae4 - 12/20/2020 09:34 AM - nobu (Nobuyoshi Nakada)
win32: Added rb_w32_ureaddir only for UTF-8 [Feature #12654]

That's not really needed. For example File.read works with any console's codepage. But Dir.entries and Dir.pwd works only for ANSI paths no matter what console's codepage is set to. There's quite inconsistency between what encodings are used and IMO best solution is just use UTF-8 everywhere.

Anyway this patch is for Ruby 3

That's not really needed. For example File.read works with any console's codepage. But Dir.entries and Dir.pwd works only for ANSI paths no matter what console's codepage is set to. There's quite inconsistency between what encodings are used and IMO best solution is just use UTF-8 everywhere.

I think they should be the console's codepage (or "locale" encoding), not UTF-8 always.
I strongly disagree. WinAPI, PowerShell and cmd supports Unicode independently of used codepage, you can navigate to paths which can't be represented with active codepage. There's really no reason to make such arbitrary limitation. Such limitation would force everyone to use UTF-8 codepage because otherwise Ruby applications won't be able to handle Unicode paths/filenames.

By default cmd opens in OEM codepage and it needs to be specifically changed. Also for example if other applications start Ruby's process with CREATE_NO_WINDOW passed to CreateProcess then Ruby will have OEM codepage or if with DETACHED_PROCESS then it will be ANSI codepage and this isn't easily changeable by parent process.

Codepages are legacy thing and it would cause only more problems and confusion. By using UTF-8 we get full Unicode support and it doesn't matter what is active codepage.

#5 - 08/05/2016 08:55 AM - usa (Usaku NAKAMURA)

Premises:

1. We don't introduce such breakage of compatibility until Ruby 3.
2. At Ruby 3, on Windows, we're planning to use UTF-8 as the default locale.
3. Ruby 3 will not force users to use UTF-8. Users will be able to choose encoding which they want to use.

The point of the issue is that users cannot choose filesystem encoding. If filesystem encoding is fixed to UTF-8, it causes other (but similar) problems.

Using locale as filesystem encoding has an advantage. Users can change locale with -E option. Then, I vote +1 to nobu's opinion.

#6 - 08/05/2016 03:25 PM - davispuh (Dāvis Mosāns)

Usaku NAKAMURA wrote:

If filesystem encoding is fixed to UTF-8, it causes other (but similar) problems.

UTF-8 can be easily encoded to any other encoding but opposite isn't always true.

But yeah I agree with other points.

#7 - 12/26/2019 05:57 AM - naruse (Yui NARUSE)

- Target version set to 3.0
- Assignee set to cruby-windows

#8 - 12/20/2020 09:34 AM - nobu (Nobuyoshi Nakada)

- Status changed from Open to Closed

Applied in changeset git|5b98b2ce39ed979aec614365a2dc3e1c30052bca.

- win32: Use UTF-8 as filesystem encoding [Feature #12654]

Co-Authored-By: Dāvis Mosāns davispuh@gmail.com

Files

<table>
<thead>
<tr>
<th>Patch Name</th>
<th>Size</th>
<th>Date</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001-On-Windows-use-UTF-8-asfilesystem-encoding.patch</td>
<td>10.4 KB</td>
<td>08/04/2016</td>
<td>davispuh (Dāvis Mosāns)</td>
</tr>
</tbody>
</table>