Remove traces of 6-byte UTF-8

04/24/2015 10:11 AM - duerst (Martin Dürst)

Status: Closed
Priority: Normal
Assignee: 
Target version: 

Description

UTF-8 was originally defined with a codespace up to 31 bits, and therefore with up to 6 bytes per character. Since quite a few years ago, it has been reduced in all the relevant definitions (ISO, Unicode, IETF) to a codespace up to 0x10FFFF and a maximum of 4 bytes per character. Many places in the Ruby code base are updated to this 4 byte limit (e.g. EncLen_UTF8 in enc/utf_8.c). But other places in the Ruby code base are not yet updated to this limit (e.g. code_to_mbclen in enc/utf_8.c). This should be fixed.

[I have classified this as a feature because I wasn't able to find a way to expose this problem in Ruby code, but this should be reclassified as a bug if such a problem can be found.]

Related issues:
- Related to Ruby trunk - Bug #13353: Backport stringio fixes  - Closed
- Related to Ruby trunk - Bug #13590: Change max byte length of UTF-8 to 4 byte...  - Closed
- Related to Ruby trunk - Feature #13588: Add Encoding#min_char_size, #max_char...  - Feedback

Associated revisions

Revision 50392 - 04/25/2015 10:36 PM - nobu (Nobuyoshi Nakada)
enc/utf_8.c: limit UTF-8

- enc/utf_8.c (code_to_mbc, code_to_mbc): reject values larger than UTF-8 max codepoints. [Feature #11094]

Revision 50392 - 04/25/2015 10:36 PM - nobu (Nobuyoshi Nakada)
enc/utf_8.c: limit UTF-8

- enc/utf_8.c (code_to_mbc, code_to_mbc): reject values larger than UTF-8 max codepoints. [Feature #11094]

Revision 50392 - 04/25/2015 10:36 PM - nobu (Nobuyoshi Nakada)
enc/utf_8.c: limit UTF-8

- enc/utf_8.c (code_to_mbc, code_to_mbc): reject values larger than UTF-8 max codepoints. [Feature #11094]

History

#1 - 04/25/2015 04:42 AM - nobu (Nobuyoshi Nakada)
- File 0001-enc-utf_8.c-pack.c-limit-UTF-8.patch added

And pack("U") and unpack("U")?

Also rubyspec seems to fail.

Array#pack with format 'U' encodes values larger than UTF-8 max codepoints ERROR
RangeError: pack(U): value out of range

11/29/2017 1/2
enc/utf_8.c: limit UTF-8

- enc/utf_8.c (code_to_mbc.len, code_to_mbc): reject values larger than UTF-8 max codepoints. [Feature #11094]

---

#3 - 03/22/2017 06:14 AM - nobu (Nobuyoshi Nakada)
- Related to Bug #13353: Backport stringio fixes added

#4 - 05/23/2017 09:11 AM - duerst (Martin Dürst)
- Related to Bug #13590: Change max byte length of UTF-8 to 4 bytes to conform to definition of UTF-8 added

#5 - 05/23/2017 09:12 AM - duerst (Martin Dürst)
- Related to Feature #13588: Add Encoding#min_char_size, #max_char_size, #minmax_char_size added

Files

<table>
<thead>
<tr>
<th>Files</th>
<th>Size</th>
<th>Date</th>
<th>Author</th>
</tr>
</thead>
</table>