Ruby master - Bug #15908

Detecting BOM with non-UTF encoding

06/08/2019 12:44 PM - nobu (Nobuyoshi Nakada)

Status:	Closed		
Priority:	Normal		
Assignee:			
Target version:			
ruby -v:		**Backport:**	2.4: UNKNOWN, 2.5: UNKNOWN, 2.6: UNKNOWN

Description

Currently, "bom|" encoding prefix to File.open is ignored if the encoding name is not a UTF.
But one usage of BOM is to tell if the stream is a UTF or not, and especially common on Windows, e.g. UTF-16LE or OEMCP.
So I think this restriction should be removed.

Related issues:

Related to Ruby master - Bug #15210: UTF-8 BOM should be removed from String ...	**Closed**

History

#1 - 06/13/2019 10:02 AM - nobu (Nobuyoshi Nakada)

- Related to Bug #15210: UTF-8 BOM should be removed from String in internal representation added

#2 - 08/29/2019 06:50 AM - duerst (Martin Dürst)

- Status changed from Open to Closed

Depending on usage, distinction of UTF-8 (with/without BOM), UTF-16LE without BOM, UTF-16BE with or without BOM, and so on may also be necessary. Also, for Japanese, traditionally distinction between EUC-JP, Shift_JIS, and ISO-2022-JP can additionally be necessary.

For more complex cases, heuristics are needed. On the other hand, applications may not want to (or not be allowed to, as e.g. for the bootstrap phase of an XML parser) allow more than a well defined subset.

This kind of processing is therefore better left to applications.

I'm closing this issue to not leave it dangling, but please feel free to reopen if you disagree.

#3 - 08/29/2019 06:54 AM - naruse (Yui NARUSE)

I understand there's theoretically exist a situation this feature is useful.
But I think it doesn't exist in practice.
I object to provide an additional utility to support legacy encoding.

#4 - 08/30/2019 02:46 AM - nobu (Nobuyoshi Nakada)

I thought UTF-16LE and CP932 as the main purpose however, I'm bit surprised that these texts have been extinct on Windows already. :tada:

#5 - 08/30/2019 08:07 AM - duerst (Martin Dürst)

nobu (Nobuyoshi Nakada) wrote:

> I thought UTF-16LE and CP932 as the main purpose however, I'm bit surprised that these texts have been extinct on Windows already. :tada:

They are not yet extinct, unfortunately :-(. In Japan, there may be quite a few cases where this would work, but even in Japan, there are many other cases where a larger and/or different selection of encodings is needed.

Files

0001-Enable-BOM-detection-with-non-UTF-encodings.patch	4.27 KB	06/08/2019	nobu (Nobuyoshi Nakada)