Feedback related to the relationship of Unicode encoding schemes and UTF encodings in the Web Platform

Henri Sivonen (hsivonen@mozilla.com)

2018-05-02

In a couple places, Unicode 10.0 makes informative statements about Encoding Schemes that the reader might mistakenly believe to be also true of how encodings work on the Web. It would be useful to add informative statements pointing out how the Web Platform behaves differently.

On the page 41 in section *2.6 Encoding Schemes*, there is a paragraph that starts with the sentence “The Internet Assigned Numbers Authority (IANA) maintains a registry of *charset names* used on the Internet.” and then goes on to mention, among other things, “some important differences may arise in terms of the requirements for each, particularly when it comes to handling of the byte order mark”.

Considering how prominent the Web Platform is, It would be worthwhile to add another paragraph along the lines of:

> Since Web browsers have deviated from what has been codified in the IANA registry, the WHATWG Encoding Standard[1] specifies a Web-compatible mapping from *labels* (similar to IANA *charset names*) to *encodings*, which differ from the *encoding schemes* in this Standard by byte order mark handling not being defined as part of the encodings themselves but instead being factored into wrapper algorithms. The WHATWG Encoding Standard requires authors to use the UTF-8 encoding (which, when used as part of the “UTF-8 decode” wrapper algorithm, matches the UTF-8 encoding scheme defined in this Standard combined with the preferred error handling policy presented in the subsection *Best Practices for Using U+FFFD* of *Section 3.9 Unicode Encoding Forms*), specifies Web-compatible handling of the byte order mark as part of wrapper algorithms and specifies two UTF-16-based encodings: UTF-16LE and UTF-BE. Since the WHATWG Encoding Standard factors byte order mark handling as a separate concern from the encodings themselves while this Standard includes byte order mark handling as part of the encoding schemes, the UTF-16LE and UTF-16BE *encodings* are not the same as the UTF-16LE and UTF-16BE *encoding schemes*. In particular, the “decode” wrapper algorithm combines the UTF-16LE and UTF-16BE *encodings* with byte order mark handling whereas the UTF-16LE and UTF-16BE *encoding schemes* explicitly do not involve byte order mark handing. Additionally, “`utf-16`” is defined as a label of the UTF-16LE encoding making little-endian treatment the default in the absence of a byte order mark for

1 https://encoding.spec.whatwg.org/

> Web content labeled “`utf-16`” on the HTTP layer. UTF-32-based encoding schemes or encodings are not supported by the Web Platform.

On page 132 in section *3.10 Unicode Encoding Schemes*, D96, D97 and D98 describe behavior that doesn't match the behavior of the Web Platform, which may cause confusion. It would be useful to add an informative note providing a cross-reference to the text proposed above for an explanation of how the UTF-16-based *encodings* in the Web Platform in combination with the “decode” algorithm from the WHATWG Encoding Standard differ from the UTF-16-based Unicode *encoding schemes* and how the Web Platform does not support UTF-32-based encodings or encoding schemes.