L2/22-110R

Addressing inconsistencies in UAX #31

To: UTC
From: Robin Leroy, Mark Davis, Source code ad hoc working group
Date: 2022-07-21

While working on UAX #31, the source code ad hoc working group noticed some inconsistencies in Unicode Standard Annex #31 *Unicode Identifier and Pattern Syntax*. These have been called out by review notes in revision 36, draft 5 of the annex for Unicode 15.0β.

This document proposes changes to UAX #31 to address these inconsistencies.

I. Proposed changes to Section 2.3 *Layout and Format Control Characters*.

Revision 36, draft 5 has the following paragraph and review note.

> Variation selectors, in particular, including standardized variants and sequences from the Ideographic Variation Database, are not included in the default identifier syntax. These are subject to the same considerations as for other Default_Ignorable_Code_Points listed above. Because variation selectors request a difference in display but do not guarantee it, they do not work well in general-purpose identifiers. The NFKC_Casefold operation can be used to remove them, along with other Default_Ignorable_Code_Points. However, in some environments it may be useful to retain variation sequences in the display form for identifiers. For more information, see *Section 1.3, Display Format*.
>
> Review Note: The sentence "Variation selectors [...] are not included in the default identifier syntax" is incorrect: The variation selectors, as well as other default ignorable code points, are part of XID_Continue.

Proposal: Change the paragraph above the review note, and add a paragraph mentioning the General Security Profile, as follows.

> While not all Default_Ignorable_Code_Points are in XID_Continue, the variation selectors *are* included in XID_Continue. These variation selectors are used in standardized variation sequences, sequences from the Ideographic Variation Database, and emoji variation sequences. However, they are subject to the same considerations as for other Default_Ignorable_Code_Points listed above. Because variation selectors request a difference in display but do not guarantee it, they do not work well in general-purpose identifiers. A profile should be used to remove them from general-purpose identifiers (along with other Default_Ignorable_Code_Points), unless their use is required in a particular domain, such as in a profile that includes emoji. For such a profile it may be useful to explicitly retain or even add certain Default_Ignorable_Code_Points in the identifier syntax.
>
> In any environment where the display form for identifiers differs from the form used to compare them, Default_Ignorable_Code_Points should be ignored for comparison. For example, this applies to case-insensitive identifiers, and in particular for any implementation that uses the

> NFKC_Casefold operation, which ignores Default_Ignorable_Code_Points. For more information, see Section 1.3, Display Format.
>
> The General Security Profile defined in Section 3.1, General Security Profile for Identifiers, in *UTS #39, Unicode Security Mechanisms* [UTS39], excludes all Default_Ignorable_Code_Points by default, including variation selectors.

II. Proposed changes to UAX31-R7 *Filtered Case-Insensitive Identifiers*.

Revision 36, draft 5 has the following review note.

> Review Note: The last sentence of this requirement incorrectly refers to Normalization Form. It should read "Except for identifiers containing excluded characters, allowed identifiers must be in the specified case folded form~~ Normalization Form~~".

Proposal: Change the requirement according to this review note, as follows:

> ***UAX31-R7**. **Filtered Case-Insensitive Identifiers:*** To meet this requirement, an implementation shall specify either simple or full case folding, and adhere to the Unicode specification for that folding. Except for identifiers containing excluded characters, allowed identifiers must be in the specified case folded form~~ Normalization Form~~.

III. Proposed changes to the note in UAX31-R7.

Revision 36, draft 5 has the following review note.

> Review Note: \P{isCasefolded} is the wrong set to disallow, as that disallows neither case, but disallows numbers. It is the set \p{Changes_When_Casefolded} that should be disallowed.

Proposal: change the note as follows:

> **Note:** For requirement UAX31-R7 with full case folding, filtering involves disallowing any characters in the set \p{Changes_When_Casefolded}~~\P{isCasefolded}~~.