Addressing inconsistencies in UAX #31

To: UTC
From: Robin Leroy, Mark Davis, Source code ad hoc working group
Date: 2022-06-09

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 review note.

> 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.

> 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~~ A profile can be used to remove them, along with other Default_Ignorable_Code_Points. However, in some environments, such as in a profile that includes emoji, it may be useful to retain some or all Default_Ignorable_Code_Points in the identifier syntax.~~variation sequences in the display form for identifiers.~~
>
> In environments where the display form for identifiers differs from the form used to compare them, as is the case for case-insensitive identifiers, Default_Ignorable_Code_Points should be ignored for comparison. For example, an implementation of case-insensitive and equivalent normalized identifiers which uses the NFKC_Casefold operation for comparison 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 [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}~~.