L2/13-169

Title: Inconsistencies between "@missing" directives in UCD files
Source: Ken Whistler (SAP AG) and Laurențiu Iancu (Microsoft Corporation)
Status: Individual contribution
Action: For consideration by the Unicode Technical Committee
Date: 2013-07-29

1. Abstract

In the UCD files, the specially formatted "@missing" comment lines are directives that provide default property values in a machine-readable form. The "@missing" directives are not consistent across the UCD files that use them. They are also not properly specified. The inconsistencies and the lack of a complete specification lead to UCD maintenance difficulties and attract error reports from users. This document compares the various patterns employed for "@missing" lines in different UCD files and proposes changes to resolve the inconsistencies between them.

2. Variety of "@missing" directives

Default property values are defined in Definition D26 of the Unicode core specification to be values of character properties given to unassigned code points or to assigned characters for which those properties are irrelevant [1]. One way to specify default property values in the UCD files is by specially formatted comment lines using the keyword "@missing" [2].

Because the "@missing" directives follow the patterns of the lines in the UCD files where they are present, and since the files do not all have the same field patterns per line, the "@missing" directives also have different formats. However, the "@missing" directives are intended to be machine readable and, to be parsed with ease, should follow consistent conventions across UCD files. The paragraphs below summarize the differences between "@missing" directives found in the various UCD files. A complete listing of the "@missing" lines is given in the Appendix.

Number and type of fields per line. In the UCD 6.3.0 files as of July 2013, the "@missing" directives use five different conventions:

Convention 1: `# @missing: code_point_range; dflt_prop_val_1`
where `dflt_prop_val_1` is an actual value or a placeholder tag such as <none> or <script>
Used in most primary and derived UCD files: BidiMirroring.txt, Blocks.txt, etc.

E.g.: `# @missing: 0000..10FFFF; <none>`

Convention 2: `# @missing: code_point_range; dflt_prop_val_1; dflt_prop_val_2`
Used only in BidiBrackets.txt (introduced in Unicode 6.3; in draft stage as of July 2013)

E.g.: `# @missing: 0000..10FFFF; <none>; n`

Convention 3: `# @missing: code_point_range; dflt_val_1; dflt_val_2; dflt_val_3`
Used in SpecialCasing.txt (with an extra semicolon) and extracted\DerivedNumericValues.txt

E.g.: `# @missing: 0000..10FFFF; <slc>; <stc>; <suc>;`

Convention 4: `# @missing: code_point_range; property_name; dflt_prop_val`
where `property_name` may be either a long or an abbreviated property name alias
and `dflt_prop_val` is, again, an actual value or a placeholder tag such as `<code point>`
Used in PropertyValueAliases.txt and DerivedNormalizationProps.txt

E.g.: `# @missing: 0000..10FFFF; NFD_QC; Yes`

Convention 5: `# @missing: code_point_range; status; dflt_prop_val`
where `status` is a special field that distinguishes between types of case folding
and `dflt_prop_val` is the placeholder tag `<code point>`
Used in CaseFolding.txt

E.g.: `# @missing: 0000..10FFFF; C; <code point>`

Note that certain UCD files, such as UnicodeData.txt or ArabicShaping.txt, do not include "`@missing`" directives. The reason is that either they are not applicable for the legacy file format (UnicodeData.txt) or that they would bloat the file (ArabicShaping.txt) because the default `Joining_Type` property values depend on the `General_Category` property values. The UCD files that are exceptional in that respect are documented in UAX #44 [2].

Placeholder tags used in "`@missing`" lines. The default property values given in "`@missing`" lines are either actual property value aliases or placeholder tags, which denote values to be substituted in for a given code point. The tags in current use and their meaning are the following:

`<none>`	the empty string
`<code point>`	the string representation of the code point value
`<script>`	the value of the catalog property `Script` (`sc`)
`<slc>`, `<stc>`, `<suc>`	the values of the string properties `Simple_Lowercase_Mapping` (`slc`), `Simple_Titlecase_Mapping` (`stc`), and `Simple_Uppercase_Mapping` (`suc`)

The tags are also inconsistent: <script> is neither a long property alias, which would be capitalized, nor an abbreviated property alias, as is used in the last three tags.

Multiple "`@missing`" lines for the same properties. Some properties are specified without "`@missing`" lines for their default values; other properties have "`@missing`" lines in either their respective UCD files or in PropertyValueAliases.txt, but not both; and other properties have "`@missing`" lines in both their respective UCD files and PropertyValueAliases.txt. Examples of each category include the following:

Neither file	White_Space or any other binary property: no "`@missing`" directives
Respective UCD file	Block: "`@missing`" line in Blocks.txt but not PropertyValueAliases.txt
PropertyValueAliases.txt	Jamo_Short_Name: "`@missing`" line in PropertyValueAliases.txt but not Jamo.txt
Both files	Bidi_Mirroring_Glyph: "`@missing`" lines in both BidiMirroring.txt and PropertyValueAliases.txt

The presence of "@missing" directives for the same property in two files creates an interdependency that adds to the maintenance of the UCD.

3. Documentation

UAX #44 succinctly describes the "@missing" directives in *Section 4.2.9, Default Values*. A more complete specification of the format of "@missing" directives seems desirable to be documented in UAX #44, especially if the UTC decides to define them more formally.

4. Error reports

The inconsistencies between "@missing" directives have attracted error reports from users. Here are some data points for reference:

July 2009 — During the Unicode 5.2 beta review period, an error report was submitted about the "@missing" directive in DerivedNormalizationProps.txt. The report noted that UCD files with multiple properties should use an "@missing" directive format that includes the property name (as in convention 4 above) to avoid ambiguities in interpreting the default value <code point>. The fix was to change

```
# @missing: 0000..10FFFF; <code point>
```

to

```
# @missing: 0000..10FFFF; NFKC_CF; <code point>
```

June 2012 — An erratum was issued for the "@missing" line in extracted\DerivedNumericValues.txt in Unicode 6.1 (and incorporated in Unicode 6.2), changing

```
# @missing: 0000..10FFFF; ; NaN
```

to

```
# @missing: 0000..10FFFF; NaN; ; NaN
```

which follows the line format of the file,

```
0F33          ; -0.5 ; ; -1/2 # No       TIBETAN DIGIT HALF ZERO
```

April 2013 — Feedback on PRI #232, Proposed Update UAX #9, Unicode Bidirectional Algorithm, noted the inconsistency between the "@missing" lines in BidiBrackets.txt and PropertyValueAliases.txt, and the lack of an "@missing" directive in the latter file for the Bidi_Paired_Bracket_Type property.

The April 2013 report referred to BidiBrackets.txt, but addressing it needs to take into account the broader problem of "@missing" directives across the UCD.

5. Solution options

The issues with the "@missing" directives can be addressed in a number of ways, as summarized below in increasing order of extent of changes.

Option 1: Do nothing. The "`@missing`" directive in BidiBrackets.txt follows the format of the data lines in that file:

```
# @missing: 0000..10FFFF; <none>; n
0028; 0029; o # LEFT PARENTHESIS
```

The same applies to other existing UCD files. One can be parsed in a similar way as the other. Also, BidiBrackets.txt (which follows convention 2) is not the only UCD file with multiple default property value fields per line – cf. SpecialCasing.txt and DerivedNumericValues.txt (convention 3).

Option 1a: Additionally to option 1, document the "`@missing`" format more thoroughly in UAX #44.

Option 2: Discard the "`@missing`" directive from BidiBrackets.txt altogether and instead document the default values for BidiBrackets.txt in PropertyValueAliases.txt, by adding an "`@missing`" directive for the Bidi_Paired_Bracket_Type property, reading:

```
# @missing: 0000..10FFFF; Bidi_Paired_Bracket_Type; n
```

This option would eliminate the inconsistency between BidiBrackets.txt and PropertyValueAliases.txt as well as avoid introducing one more instance of multivalued "`@missing`" directives, which are few in the UCD.

Option 2a: Additionally to option 2, document the default values of `Bidi_Paired_Bracket` and `Bidi_Paired_Bracket_Type` in UAX #44.

Option 3: Delete all of the multivalued "`@missing`" directives (conventions 2 and 3; potentially 5 as well) and add corresponding single-valued directives to PropertyValueAliases.txt. The advantage of this approach would be that only two "`@missing`" formats would remain: one with a single default property value (convention 1) and the other with a property name and a default property value (convention 4), which are also the dominant conventions.

Option 4: A generalization of option 3 would be to attempt to make the "`@missing`" convention as consistent as possible, using the PropertyValueAliases.txt file as the model and the ordinary home for it. The advantage is greatest consistency, but the disadvantage is biggest hit, requiring touching more than 20 UCD files, both primary and derived. It might also introduce new representation problems for the unusual formats in SpecialCasing.txt (convention 3) and CaseFolding.txt (convention 5).

Necessary changes. Regardless of the solution option chosen, adding an "`@missing`" directive in PropertyValueAliases.txt for the `Bidi_Paired_Bracket_Type` property seems adequate and it would also address the respective part of the PRI #232 feedback.

6. References

[1] The Unicode Consortium. *The Unicode Standard, Version 6.3.0*, (Mountain View, CA: The Unicode Consortium, 2013. ISBN 978-1-936213-08-5). http://www.unicode.org/versions/Unicode6.3.0/.

[2] Proposed Update Unicode Standard Annex #44, *Unicode Character Database*, edited by Mark Davis, Laurențiu Iancu and Ken Whistler, an integral part of *The Unicode Standard*. Version 6.3.0 (draft 8), 2013-07-26. (http://www.unicode.org/reports/tr44/tr44-11d8.html). Latest Version: http://www.unicode.org/reports/tr44/.

Appendix: Complete listing of "@missing" lines in the UCD 6.3 files as of July 2013

A complete listing of "@missing" lines in the UCD 6.3 files as of July 2013 is given below.

Convention 1 (one default property value):

```
BidiMirroring-6.3.0d2.txt:51:# @missing: 0000..10FFFF; <none>
Blocks-6.3.0d1.txt:28:# @missing: 0000..10FFFF; No_Block
DerivedAge-6.3.0d11.txt:47:# @missing: 0000..10FFFF; Unassigned
EastAsianWidth-6.3.0d2.txt:39:# @missing: 0000..10FFFF; N
HangulSyllableType-6.3.0d1.txt:16:# @missing: 0000..10FFFF; Not_Applicable
IndicMatraCategory-6.3.0d1.txt:85:# @missing: 0000..10FFFF; NA
IndicSyllabicCategory-6.3.0d1.txt:60:# @missing: 0000..10FFFF; Other
LineBreak-6.3.0d3.txt:49:# @missing: 0000..10FFFF; XX
ScriptExtensions-6.3.0d11.txt:18:# @missing: 0000..10FFFF; <script>
Scripts-6.3.0d19.txt:16:# @missing: 0000..10FFFF; Unknown
auxiliary\GraphemeBreakProperty-6.3.0d11.txt:16:# @missing: 0000..10FFFF; Other
auxiliary\SentenceBreakProperty-6.3.0d19.txt:16:# @missing: 0000..10FFFF; Other
auxiliary\WordBreakProperty-6.3.0d19.txt:16:# @missing: 0000..10FFFF; Other
extracted\DerivedBidiClass-6.3.0d19.txt:64:# @missing: 0000..10FFFF; Left_To_Right
extracted\DerivedCombiningClass-6.3.0d19.txt:16:# @missing: 0000..10FFFF; Not_Reordered
extracted\DerivedDecompositionType-6.3.0d1.txt:16:# @missing: 0000..10FFFF; None
extracted\DerivedEastAsianWidth-6.3.0d19.txt:16:# @missing: 0000..10FFFF; Neutral
extracted\DerivedJoiningGroup-6.3.0d1.txt:16:# @missing: 0000..10FFFF; No_Joining_Group
extracted\DerivedJoiningType-6.3.0d11.txt:16:# @missing: 0000..10FFFF; Non_Joining
extracted\DerivedLineBreak-6.3.0d19.txt:16:# @missing: 0000..10FFFF; Unknown
extracted\DerivedNumericType-6.3.0d1.txt:24:# @missing: 0000..10FFFF; None
```

Convention 2 (two default property values):

```
BidiBrackets-6.3.0d3.txt:51:# @missing: 0000..10FFFF; <none>; n
```

Convention 3 (three default property values or fields):

```
SpecialCasing-6.3.0d18.txt:51:# @missing: 0000..10FFFF; <slc>; <stc>; <suc>;
extracted\DerivedNumericValues-6.3.0d1.txt:32:# @missing: 0000..10FFFF; NaN; ; NaN
```

Convention 4 (property name and default property value):

```
DerivedNormalizationProps-6.3.0d11.txt:742:# @missing: 0000..10FFFF; NFD_QC; Yes
DerivedNormalizationProps-6.3.0d11.txt:996:# @missing: 0000..10FFFF; NFC_QC; Yes
DerivedNormalizationProps-6.3.0d11.txt:1129:# @missing: 0000..10FFFF; NFKD_QC; Yes
DerivedNormalizationProps-6.3.0d11.txt:1680:# @missing: 0000..10FFFF; NFKC_QC; Yes
DerivedNormalizationProps-6.3.0d11.txt:2852:# @missing: 0000..10FFFF; NFKC_CF; <code point>
PropertyValueAliases-6.3.0d18.txt:124:# @missing: 0000..10FFFF; Bidi_Mirroring_Glyph; <none>
PropertyValueAliases-6.3.0d18.txt:128:# @missing: 0000..10FFFF; Bidi_Paired_Bracket; <none>
PropertyValueAliases-6.3.0d18.txt:422:# @missing: 0000..10FFFF; Case_Folding; <code point>
PropertyValueAliases-6.3.0d18.txt:476:# @missing: 0000..10FFFF; Decomposition_Mapping;
                                                                    <code point>
PropertyValueAliases-6.3.0d18.txt:550:# @missing: 0000..10FFFF; FC_NFKC_Closure; <code point>
PropertyValueAliases-6.3.0d18.txt:670:# @missing: 0000..10FFFF; ISO_Comment; <none>
PropertyValueAliases-6.3.0d18.txt:721:# @missing: 0000..10FFFF; Jamo_Short_Name; <none>
PropertyValueAliases-6.3.0d18.txt:921:# @missing: 0000..10FFFF; NFKC_Casefold; <code point>
```

```
PropertyValueAliases-6.3.0d18.txt:936:# @missing: 0000..10FFFF; Name; <none>
PropertyValueAliases-6.3.0d18.txt:940:# @missing: 0000..10FFFF; Name_Alias; <none>
PropertyValueAliases-6.3.0d18.txt:956:# @missing: 0000..10FFFF; Numeric_Value; NaN
PropertyValueAliases-6.3.0d18.txt:1132:# @missing: 0000..10FFFF; Script_Extensions; <script>
PropertyValueAliases-6.3.0d18.txt:1154:# @missing: 0000..10FFFF; Simple_Case_Folding;
                                                                  <code point>
PropertyValueAliases-6.3.0d18.txt:1158:# @missing: 0000..10FFFF; Simple_Lowercase_Mapping;
                                                                  <code point>
PropertyValueAliases-6.3.0d18.txt:1162:# @missing: 0000..10FFFF; Simple_Titlecase_Mapping;
                                                                  <code point>
PropertyValueAliases-6.3.0d18.txt:1166:# @missing: 0000..10FFFF; Simple_Uppercase_Mapping;
                                                                  <code point>
PropertyValueAliases-6.3.0d18.txt:1180:# @missing: 0000..10FFFF; Unicode_1_Name; <none>
PropertyValueAliases-6.3.0d18.txt:1234:# @missing: 0000..10FFFF; cjkAccountingNumeric; NaN
PropertyValueAliases-6.3.0d18.txt:1238:# @missing: 0000..10FFFF; cjkCompatibilityVariant;
                                                                  <code point>
PropertyValueAliases-6.3.0d18.txt:1242:# @missing: 0000..10FFFF; cjkIICore; <none>
PropertyValueAliases-6.3.0d18.txt:1246:# @missing: 0000..10FFFF; cjkIRG_GSource; <none>
PropertyValueAliases-6.3.0d18.txt:1250:# @missing: 0000..10FFFF; cjkIRG_HSource; <none>
PropertyValueAliases-6.3.0d18.txt:1254:# @missing: 0000..10FFFF; cjkIRG_JSource; <none>
PropertyValueAliases-6.3.0d18.txt:1258:# @missing: 0000..10FFFF; cjkIRG_KPSource; <none>
PropertyValueAliases-6.3.0d18.txt:1262:# @missing: 0000..10FFFF; cjkIRG_KSource; <none>
PropertyValueAliases-6.3.0d18.txt:1266:# @missing: 0000..10FFFF; cjkIRG_MSource; <none>
PropertyValueAliases-6.3.0d18.txt:1270:# @missing: 0000..10FFFF; cjkIRG_TSource; <none>
PropertyValueAliases-6.3.0d18.txt:1274:# @missing: 0000..10FFFF; cjkIRG_USource; <none>
PropertyValueAliases-6.3.0d18.txt:1278:# @missing: 0000..10FFFF; cjkIRG_VSource; <none>
PropertyValueAliases-6.3.0d18.txt:1282:# @missing: 0000..10FFFF; cjkOtherNumeric; NaN
PropertyValueAliases-6.3.0d18.txt:1286:# @missing: 0000..10FFFF; cjkPrimaryNumeric; NaN
PropertyValueAliases-6.3.0d18.txt:1290:# @missing: 0000..10FFFF; cjkRSUnicode; <none>
```

Convention 5 (status field and default property value):

```
CaseFolding-6.3.0d1.txt:61:# @missing: 0000..10FFFF; C; <code point>
```