Re:	Default Ignorable Issues
From:	Mark Davis, Ken Whistler
Date:	2002-10-30

The following is for action 92-A52:

> *Mark Davis, Ken Whistler; Clarify the situation regarding the distinctions among characters that have no visible glyph for the November 2002 UTC. [L2/02-267R3]*

This document contains suggested text, the UTC issues that have arisen, plus some background information.

Suggested Text

Default-ignorable code points are those that should be ignored by default in rendering (unless explicitly supported). They have no visible glyph or advance width in and of themselves, although they may affect the display, positioning, or adornment of adjacent or surrounding characters. Some of the default ignorable code points are assigned characters, while others are reserved for future assignment.

An implementation should ignore default-ignorable characters in rendering whenever it does *not* support the characters.

This can be contrasted with the situation for non-default-ignorable characters. If an implementation does not support, say, a U+0915 (क) DEVANAGARI LETTER KA, it should still not ignore it in rendering. Displaying *nothing* would give the user the impression that it does not occur in the text at all. The recommendation in that case is to display a "last-resort" glyph or a visible "missing glyph" box; see Section 5.3 Unknown and Missing Characters.

With default-ignorable characters, such as U+200D (

) ZERO WIDTH JOINER, the

situation is different. If the program does not support that character, best practice is to ignore it completely -- without displaying a last-resort glyph or a visible box -- since the normal display of the character is invisible: its effects are on other characters. And since the character is not supported, those effects cannot be shown.

Other characters will have other effects on adjacent characters. For example:

- U+070F (

) SYRIAC ABBREVIATION MARK is an example of a character that has no visible form, unless it is followed by letters (and optionally, combining marks). In the latter circumstance, the character causes those letters to to be adorned by an overline, as described in Section 8.3 Syriac.
- U+2060 (
) WORD JOINER is an example of a character that doesn't produce a visible change in the appearance of surrounding characters; instead, its only effect is to indicated that there should be no line break at that point.

- U+2061 ( ) FUNCTION APPLICATION has no effect at all on display, and is only used in internal mathematical expression processing.
- U+00AD ( ) SOFT HYPHEN has a null default appearance: the appearance of "ther

  apist" is simply "therapist"; no visible glyph. In linebreak processing, it indicates a possible intra-word break. At any intra-word break that is used for a line break -- whether resulting from this character or by automatic process -- a hyphen glyph (perhaps with spelling changes) or some other indication can be shown, depending on language and context.

This does *not* imply that default-ignorable code points must always be invisible: an implementation can show a visible glyph on request, such as in a "Show Hidden" mode. A particular use of a "Show Hidden" mode is to show a visible indication of "misplaced" or "ineffectual" formatting codes. For example, this would include two adjacent U+200D ( ) ZERO WIDTH JOINER characters, where the extra character has no effect at all.

The default-ignorable *unassigned* code points lie in particular designated ranges. These ranges are designed and reserved for future default-ignorable characters, to allow forward compatibility. All implementations should ignore all unassigned default ignorable code points in all rendering. Any new default-ignorable characters should be assigned in those ranges, permitting existing programs to ignore them until they are supported in some future version of the program.

There are other characters that have no visible glyphs: the white-space characters. These typically have advance-width, however. The line separation characters such as CR do not clearly exhibit this advance-width since they are always at the end of a line, but in most GUIs show a visible advance width when selected.

UTC Issues:

Should the following be included in Default_Ignorable_Code_Points?

1. All Noncharacters
   - the recommendation is yes. This appears to be an oversight. Since noncharacters can be silently removed without constituting "modification" under the conformance rules, we certainly should recommend a null glyph display for them by default.
2. U+FFFC OBJECT REPLACEMENT CHARACTER
   - if so, there are two possible options:
     1. Make U+FFFC Other_Default_Ignorable_Code_Point, despite its general category.
     2. Change gc from So to Cf, which would automatically cause it to be a member of *Default_Ignorable_Code_Points.*
3. U+06DD ARABIC END OF AYAH
   - this is an edge case, and requires discussion. The key is whether it should be displayed as nothing if there are no suitable characters after it, or if it should be displayed as a visible glyph.
   - in other words, the issue is whether this is a "prepositive aggregator", but with a visible default glyph (see below).

Background Information

The following provides background information relevant to the discussion of the above text and issues, with a breakdown according to different properties of "characters that have no visible glyphs":

1. spaces (gc=Zs)
2. control codes (gc=Cc)
3. format control characters (gc=Cf)
4. whitespace (White_Space=true)
5. default ignorables (Default_Ignorable_Code_Point=true)
6. invisible placeholders (n/a)
7. noncharacters (Noncharacter_Code_Point=true)

```
1. Spaces (gc=Zs)

0020          ; Zs #       SPACE
00A0          ; Zs #       NO-BREAK SPACE
1680          ; Zs #       OGHAM SPACE MARK
2000..200B    ; Zs #  [12] EN QUAD..ZERO WIDTH SPACE
202F          ; Zs #       NARROW NO-BREAK SPACE
205F          ; Zs #       MEDIUM MATHEMATICAL SPACE
3000          ; Zs #       IDEOGRAPHIC SPACE

1a. Non-break spaces

00A0          ; Zs #       NO-BREAK SPACE
202F          ; Zs #       NARROW NO-BREAK SPACE

1b. Zero-width spaces

200B          ; Zs #       ZERO WIDTH SPACE

1c. Designated-width spaces (other than zero-width)

2000..200A    ; Zs #  [11] EN QUAD..HAIR SPACE
205F          ; Zs #       MEDIUM MATHEMATICAL SPACE
3000          ; Zs #       IDEOGRAPHIC SPACE

1c.i. Full-width spaces (gc=Zs and ea=F)

3000          ; Zs #       IDEOGRAPHIC SPACE

1d. Spaces that may have visible rendering

1680          ; Zs #       OGHAM SPACE MARK
```

```
2. Control codes (gc=Cc)

0000..001F    ; Cc #  [32] <control>..<control>
007F..009F    ; Cc #  [33] <control>..<control>

2a. Control codes with Unicode semantics (not default ignorable)

0009..000D    ; White_Space # Cc   [5] <control>..<control>
0085          ; White_Space # Cc       <control>

2b. Unicode null terminator (for C, C++, etc.)

0000          ; Cc #       <control>

2c. Unicode escape character (for 10646 announcement mechanisms)

001B          ; Cc #       <control>
```

```
3. Format control characters (gc=Cf)

00AD          ; Cf #       SOFT HYPHEN
06DD          ; Cf #       ARABIC END OF AYAH
070F          ; Cf #       SYRIAC ABBREVIATION MARK
180E          ; Cf #       MONGOLIAN VOWEL SEPARATOR
200C..200F    ; Cf #   [4] ZERO WIDTH NON-JOINER..RIGHT-TO-LEFT MARK
202A..202E    ; Cf #   [5] LEFT-TO-RIGHT EMBEDDING..RIGHT-TO-LEFT OVERRIDE
2060..2063    ; Cf #   [4] WORD JOINER..INVISIBLE SEPARATOR
206A..206F    ; Cf #   [6] INHIBIT SYMMETRIC SWAPPING..NOMINAL DIGIT SHAPES
FEFF          ; Cf #       ZERO WIDTH NO-BREAK SPACE
FFF9..FFFB    ; Cf #   [3] INTERLINEAR ANNOTATION ANCHOR..INTERLINEAR ANNOTATION
TERMINATOR
1D173..1D17A  ; Cf #   [8] MUSICAL SYMBOL BEGIN BEAM..MUSICAL SYMBOL END PHRASE
E0001         ; Cf #       LANGUAGE TAG
E0020..E007F  ; Cf #  [96] TAG SPACE..CANCEL TAG

3a. Bidi Control (Bidi_Control=true)

200E..200F    ; Bidi_Control # Cf   [2] LEFT-TO-RIGHT MARK..RIGHT-TO-LEFT MARK
202A..202E    ; Bidi_Control # Cf   [5] LEFT-TO-RIGHT EMBEDDING..RIGHT-TO-LEFT
OVERRIDE

3b. Join Control (Join_Control=true)

200C..200D    ; Join_Control # Cf   [2] ZERO WIDTH NON-JOINER..ZERO WIDTH JOINER

3c. Hyphenation Control

00AD          ; Cf #       SOFT HYPHEN

3d. Boundary Determination Control

2060          ; Cf #       WORD JOINER
FEFF          ; Cf #       ZERO WIDTH NO-BREAK SPACE (deprecated in this function)

3e. Byte Order Mark (signature)

FEFF          ; Cf #       ZERO WIDTH NO-BREAK SPACE

3f. Prepositive Aggregators

06DD          ; Cf #       ARABIC END OF AYAH
070F          ; Cf #       SYRIAC ABBREVIATION MARK

3g. Annotation bracketing

FFF9..FFFB    ; Cf #   [3] INTERLINEAR ANNOTATION ANCHOR..INTERLINEAR ANNOTATION
TERMINATOR

3h. Musical phrase bracketing

1D173..1D17A  ; Cf #   [8] MUSICAL SYMBOL BEGIN BEAM..MUSICAL SYMBOL END PHRASE

3i. Language tags

E0001         ; Cf #       LANGUAGE TAG
E0020..E007F  ; Cf #  [96] TAG SPACE..CANCEL TAG
```

```
4. Whitespace (White_Space=true)

0009..000D    ; White_Space # Cc   [5] <control>..<control>
0020          ; White_Space # Zs       SPACE
0085          ; White_Space # Cc       <control>
00A0          ; White_Space # Zs       NO-BREAK SPACE
1680          ; White_Space # Zs       OGHAM SPACE MARK
2000..200A    ; White_Space # Zs  [11] EN QUAD..HAIR SPACE
2028          ; White_Space # Zl       LINE SEPARATOR
2029          ; White_Space # Zp       PARAGRAPH SEPARATOR
202F          ; White_Space # Zs       NARROW NO-BREAK SPACE
205F          ; Zs          #          MEDIUM MATHEMATICAL SPACE
3000          ; White_Space # Zs       IDEOGRAPHIC SPACE
```

4a. Newlines

```
000A          ; White_Space # Cc       <control> (LF)
000B          ; White_Space # Cc       <control> (VT)
000C          ; White_Space # Cc       <control> (FF)
000D          ; White_Space # Cc       <control> (CR)
0085          ; White_Space # Cc       <control> (NEL)
2028          ; White_Space # Zl       LINE SEPARATOR
2029          ; White_Space # Zp       PARAGRAPH SEPARATOR
```

4b. Unicode separators (gc=Zl or gc=Zp)

```
2028          ; Zl #       LINE SEPARATOR
2029          ; Zp #       PARAGRAPH SEPARATOR
```

4c. Spaces with non-zero advance width

```
0020          ; White_Space # Zs       SPACE
00A0          ; White_Space # Zs       NO-BREAK SPACE
1680          ; White_Space # Zs       OGHAM SPACE MARK
2000..200A    ; White_Space # Zs  [11] EN QUAD..HAIR SPACE
202F          ; White_Space # Zs       NARROW NO-BREAK SPACE
205F          ; Zs          #          MEDIUM MATHEMATICAL SPACE
3000          ; White_Space # Zs       IDEOGRAPHIC SPACE
```

---

5. Default Ignorable Characters (derived)

```
# Derived Property: Default_Ignorable_Code_Point
#  Generated from <2060..206F, FFF0..FFFB, E0000..E0FFF>
#    + Other_Default_Ignorable_Code_Point + (Cf + Cc + Cs - White_Space)

0000..0008    ; Default_Ignorable_Code_Point # Cc   [9] <control>..<control>
000E..001F    ; Default_Ignorable_Code_Point # Cc  [18] <control>..<control>
007F..0084    ; Default_Ignorable_Code_Point # Cc   [6] <control>..<control>
0086..009F    ; Default_Ignorable_Code_Point # Cc  [26] <control>..<control>
06DD          ; Default_Ignorable_Code_Point # Cf       ARABIC END OF AYAH
070F          ; Default_Ignorable_Code_Point # Cf       SYRIAC ABBREVIATION MARK
180B..180D    ; Default_Ignorable_Code_Point # Mn   [3] MONGOLIAN FREE VARIATION
SELECTOR ONE..MONGOLIAN FREE VARIATION SELECTOR THREE
180E          ; Default_Ignorable_Code_Point # Cf       MONGOLIAN VOWEL SEPARATOR
200C..200F    ; Default_Ignorable_Code_Point # Cf   [4] ZERO WIDTH
NON-JOINER..RIGHT-TO-LEFT MARK
202A..202E    ; Default_Ignorable_Code_Point # Cf   [5] LEFT-TO-RIGHT
EMBEDDING..RIGHT-TO-LEFT OVERRIDE
2060..2063    ; Default_Ignorable_Code_Point # Cf   [4] WORD JOINER..INVISIBLE
SEPARATOR
2064..2069    ; Default_Ignorable_Code_Point # Cn   [6]
206A..206F    ; Default_Ignorable_Code_Point # Cf   [6] INHIBIT SYMMETRIC
SWAPPING..NOMINAL DIGIT SHAPES
D800..DFFF    ; Default_Ignorable_Code_Point # Cs [2048]
FE00..FE0F    ; Default_Ignorable_Code_Point # Mn  [16] VARIATION
SELECTOR-1..VARIATION SELECTOR-16
FEFF          ; Default_Ignorable_Code_Point # Cf       ZERO WIDTH NO-BREAK SPACE
FFF0..FFF8    ; Default_Ignorable_Code_Point # Cn   [9]
FFF9..FFFB    ; Default_Ignorable_Code_Point # Cf   [3] INTERLINEAR ANNOTATION
ANCHOR..INTERLINEAR ANNOTATION TERMINATOR
1D173..1D17A  ; Default_Ignorable_Code_Point # Cf   [8] MUSICAL SYMBOL BEGIN
BEAM..MUSICAL SYMBOL END PHRASE
E0000         ; Default_Ignorable_Code_Point # Cn
E0001         ; Default_Ignorable_Code_Point # Cf       LANGUAGE TAG
E0002..E001F  ; Default_Ignorable_Code_Point # Cn  [30]
E0020..E007F  ; Default_Ignorable_Code_Point # Cf  [96] TAG SPACE..CANCEL TAG
E0080..E0FFF  ; Default_Ignorable_Code_Point # Cn [3968]
```

5a. Variation selectors (invisible combining marks)

```
180B..180D    ; Default_Ignorable_Code_Point # Mn   [3] MONGOLIAN FREE VARIATION
SELECTOR ONE..MONGOLIAN FREE VARIATION SELECTOR THREE
FE00..FE0F    ; Default_Ignorable_Code_Point # Mn  [16] VARIATION
SELECTOR-1..VARIATION SELECTOR-16
```

5b. Grapheme Join Control (invisible combining mark)

```
034F          ; Mn #       COMBINING GRAPHEME JOINER

5c. Other Default Ignorable (list not captured by rule)

180B..180D    ; Other_Default_Ignorable_Code_Point # Mn   [3] MONGOLIAN FREE
VARIATION SELECTOR ONE..MONGOLIAN FREE VARIATION SELECTOR THREE
2060..2063    ; Other_Default_Ignorable_Code_Point # Cf   [4] WORD
JOINER..INVISIBLE SEPARATOR
2064..2069    ; Other_Default_Ignorable_Code_Point # Cn   [6]
206A..206F    ; Other_Default_Ignorable_Code_Point # Cf   [6] INHIBIT SYMMETRIC
SWAPPING..NOMINAL DIGIT SHAPES
FE00..FE0F    ; Other_Default_Ignorable_Code_Point # Mn  [16] VARIATION
SELECTOR-1..VARIATION SELECTOR-16
FFF0..FFF8    ; Other_Default_Ignorable_Code_Point # Cn   [9]
FFF9..FFFB    ; Other_Default_Ignorable_Code_Point # Cf   [3] INTERLINEAR
ANNOTATION ANCHOR..INTERLINEAR ANNOTATION TERMINATOR
E0000         ; Other_Default_Ignorable_Code_Point # Cn
E0001         ; Other_Default_Ignorable_Code_Point # Cf       LANGUAGE TAG
E0002..E001F  ; Other_Default_Ignorable_Code_Point # Cn  [30]
E0020..E007F  ; Other_Default_Ignorable_Code_Point # Cf  [96] TAG SPACE..CANCEL
TAG
E0080..E0FFF  ; Other_Default_Ignorable_Code_Point # Cn [3968]
```

---

6. Invisible Placeholders

```
FFFC          ; So #       OBJECT REPLACEMENT CHARACTER
```

---

7. Noncharacters

```
FDD0..FDEF    ; Noncharacter_Code_Point # Cn  [32]
FFFE..FFFF    ; Noncharacter_Code_Point # Cn   [2]
1FFFE..1FFFF  ; Noncharacter_Code_Point # Cn   [2]
2FFFE..2FFFF  ; Noncharacter_Code_Point # Cn   [2]
3FFFE..3FFFF  ; Noncharacter_Code_Point # Cn   [2]
4FFFE..4FFFF  ; Noncharacter_Code_Point # Cn   [2]
5FFFE..5FFFF  ; Noncharacter_Code_Point # Cn   [2]
6FFFE..6FFFF  ; Noncharacter_Code_Point # Cn   [2]
7FFFE..7FFFF  ; Noncharacter_Code_Point # Cn   [2]
8FFFE..8FFFF  ; Noncharacter_Code_Point # Cn   [2]
9FFFE..9FFFF  ; Noncharacter_Code_Point # Cn   [2]
AFFFE..AFFFF  ; Noncharacter_Code_Point # Cn   [2]
BFFFE..BFFFF  ; Noncharacter_Code_Point # Cn   [2]
CFFFE..CFFFF  ; Noncharacter_Code_Point # Cn   [2]
DFFFE..DFFFF  ; Noncharacter_Code_Point # Cn   [2]
EFFFE..EFFFF  ; Noncharacter_Code_Point # Cn   [2]
FFFFE..FFFFF  ; Noncharacter_Code_Point # Cn   [2]
10FFFE..10FFFF; Noncharacter_Code_Point # Cn   [2]
```