From: Peter Edberg, Mark Davis, Andy Heninger
Subject: Modifying \w (word) in UTS #18
Date: 2011-04-28

We have had the following action:

125	A095	Mark Davis,...		Write up a proposal based on Andy Heninger's feedback on UTS #18. (Re definition of [:word:])	L2/10-427

Our recommendation is to post a PRI for changing the definition of \w (word) in http://unicode.org/reports/tr18/proposed.html#Compatibility_Properties

From:

\p{alpha}
\p{gc=Mark}
\p{digit}
\p{gc=Connector_Punctuation}

To:

\p{xid_continue}

Issue: POSIX expectations may be that \w include **alnum** (**alpha+digit**). However, that doesn't appear to be in the POSIX standard. So we might not make the change if that superset relation is important. If not, we should still add Middle dot and its canonical equivalent, Ano Teleia.

Note: another approach would be to change the General Category of Middle Dot to **Lm** (Letter Modifier), which would accurately reflect its usage in Catalan. This would also help with the request from the Spanish government, and alleviate pressure to encode a letterlike Middle Dot. That would be a more extensive change, however.

The change would add the following:

The middle-dot would be welcome. The others are in xid_continue for compatibility, but wouldn't hurt.

Latin 1 Supplement — ***Latin-1 punctuation and symbols***
U+00B7 (·) MIDDLE DOT

Greek And Coptic — ***Punctuation***
U+0387 (·) GREEK ANO TELEIA

Ethiopic — ***Digits***
U+1369 ( ) ETHIOPIC DIGIT ONE
...

U+1371 ( ) ETHIOPIC DIGIT NINE

New Tai Lue — *Digits*

U+19DA ( ) NEW TAI LUE THAM DIGIT ONE

Letterlike Symbols — *Letterlike symbol*

U+2118 (℘) SCRIPT CAPITAL P
U+212E (℮) ESTIMATED SYMBOL

The change would remove the following:

We don't think any of these are necessary in words, and some are pernicious, like U+FDFA (ﷺ) ARABIC LIGATURE SALLALLAHOU ALAYHE WASALLAM

Greek And Coptic — *Iota subscript*

U+037A (ͺ) GREEK YPOGEGRAMMENI

Cyrillic — *Historic miscellaneous*

U+0488 ( ) COMBINING CYRILLIC HUNDRED THOUSANDS SIGN
U+0489 ( ) COMBINING CYRILLIC MILLIONS SIGN

Combining Diacritical Marks For Symbols — *Enclosing diacritics*

U+20DD (◌⃝) COMBINING ENCLOSING CIRCLE
U+20DE (◌⃞) COMBINING ENCLOSING SQUARE
U+20DF (◌⃟) COMBINING ENCLOSING DIAMOND
U+20E0 (◌⃠) COMBINING ENCLOSING CIRCLE BACKSLASH
U+20E2 ( ) COMBINING ENCLOSING SCREEN
U+20E3 ( ) COMBINING ENCLOSING KEYCAP
U+20E4 ( ) COMBINING ENCLOSING UPWARD POINTING TRIANGLE

Enclosed Alphanumerics — *Circled Latin letters*

U+24B6 (Ⓐ) CIRCLED LATIN CAPITAL LETTER A
…
U+24E9 (ⓩ) CIRCLED LATIN SMALL LETTER Z

Supplemental Punctuation — *Archaic punctuation*

U+2E2F ( ) VERTICAL TILDE

Cyrillic Extended B — *Combining numeric signs*

U+A670 ( ) COMBINING CYRILLIC TEN MILLIONS SIGN
U+A671 ( ) COMBINING CYRILLIC HUNDRED MILLIONS SIGN
U+A672 ( ) COMBINING CYRILLIC THOUSAND MILLIONS SIGN

Arabic Presentation Forms A — *Ligatures (two elements)*

U+FC5E (ﱞ) ARABIC LIGATURE SHADDA WITH DAMMATAN ISOLATED FORM
…

U+FC63 (ﱣ) ARABIC LIGATURE SHADDA WITH SUPERSCRIPT ALEF ISOLATED FORM

Arabic Presentation Forms A — *Word ligatures*

U+FDFA (ﷺ) ARABIC LIGATURE SALLALLAHOU ALAYHE WASALLAM
U+FDFB (ﷻ) ARABIC LIGATURE JALLAJALALOUHOU

Arabic Presentation Forms B — *Glyphs for spacing forms of Arabic points*

U+FE70 (ً) ARABIC FATHATAN ISOLATED FORM
U+FE72 (ٌ) ARABIC DAMMATAN ISOLATED FORM
U+FE74 (ٍ) ARABIC KASRATAN ISOLATED FORM
U+FE76 (َ) ARABIC FATHA ISOLATED FORM
U+FE78 (ُ) ARABIC DAMMA ISOLATED FORM
U+FE7A (ِ) ARABIC KASRA ISOLATED FORM
U+FE7C (ّ) ARABIC SHADDA ISOLATED FORM
U+FE7E (ْ) ARABIC SUKUN ISOLATED FORM