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+0087 (·) MIDDLE DOT

**Greek And Coptic** — *Punctuation*

\U+0387 (·) GREEK ANO TELEIA

**Ethiopic** — *Digits*

\U+1369 (๑) ETHIOPIC DIGIT ONE
The change would remove the following:

We don’t think any of these are necessary in words, and some are pernicious, like

\textit{Greek And Coptic — Iota subscript}

\textit{Cyrillic — Historic miscellaneous}

\textit{Combining Diacritical Marks For Symbols — Enclosing diacritics}

\textit{Enclosed Alphanumerics — Circled Latin letters}

\textit{Supplemental Punctuation — Archaic punctuation}

\textit{Cyrillic Extended B — Combining numeric signs}

\textit{Arabic Presentation Forms A — Ligatures (two elements)}
[^1] ARABIC LIGATURE SHADDA WITH SUPERSCRIPT ALEF ISOLATED FORM

**Arabic Presentation Forms A — Word ligatures**

U+FC63 ([^1]) ARABIC LIGATURE SALLALLAHOU ALAYHE WASALLAM
U+FD80 ([^2]) ARABIC LIGATURE JALLAJALALOUHOU

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

U+FE70 ([^3]) ARABIC FATHATAN ISOLATED FORM
U+FE72 ([^3]) ARABIC DAMMATAN ISOLATED FORM
U+FE74 ([^3]) ARABIC KASRATAN ISOLATED FORM
U+FE76 ([^3]) ARABIC FATHA ISOLATED FORM
U+FE78 ([^3]) ARABIC DAMMA ISOLATED FORM
U+FE7A ([^3]) ARABIC KASRA ISOLATED FORM
U+FE7C ([^3]) ARABIC SHADDA ISOLATED FORM
U+FE7E ([^3]) ARABIC SUKUN ISOLATED FORM