Handling ‘defective’ names
2005-05-10 - Asmus Freytag

L2/05-137

Unicode names are immutable – this makes it possible to refer to a character by name in a stable way, whether in documentation, user interface or as formal identifier. The restriction imposed by this policy can be described as: “Once a character is identified with a given name, it must always be possible to uniquely refer to it by that name”.

Some Unicode names are defective[†]. Others are merely not ideal, in being either non-intuitive or merely arbitrary. For some character names, British and American usage differ, creating issues for some user groups.

Recently, there’s been intense discussion of the issue of character name defects. While it is perhaps true that only a vocal minority raised the issue in a forceful way, the ensuing discussion makes clear that there is an ongoing cost to implementers and users in dealing with defective names.

Because character names are immutable, defects to character names cannot become errata. This has lead to the situation where users cannot learn about issues with character names except by somewhat unsystematic annotations in the character names list.

Characters are not the only named objects in the Unicode Standard. Properties and property values are also named. These names are also used as identifiers, in documentation, etc. and therefore must satisfy the same requirements as characters names. However property names are handled differently: If a property name is found defective (such as the block name “Cyrillic Supplementary” or the property value alias “Inseperable”) they were retained, but corrected names were added as aliases.

1) The UTC should approve formal aliases for ‘defective’ names
   Such aliases should be unique in the character name space. They should cover the typos and some or all of the ‘bad names’ section appended below.
2) The UTC should document defects or known issues with names on the web, in a manner similar to errata, but as part of a separate document. The document created by Rick & Ken would be an excellent starting poin
3) The UTC should maintain a list of names localized to American usage. This would account for differences in usage, such as CENTRE/CENTER, SOLIDUS/SLASH, LOW LINE/UNDERSCORE etc.
4) Loose name matching should be extended to make the difference between LIGATURE and LETTER ignorable.

[†] A defective name is one that is misspelled, highly misleading or both. The use of letter vs ligature does not qualify, but ‘BRAKCET’ vs. ‘BRACKET’ does. The use of LETTER O I for LETTER GHA might qualify as an example of ‘highly misleading’. The use of CENTER/CENTRE or PERIOD/FULL STOP are not defects, but localization issues.

Name issues sorted by severity

Typos

FE18;PRESENTATION FORM FOR VERTICAL RIGHT WHITE LENTICULAR BRACKET
1D0C5;BYZANTINE MUSICAL SYMBOL FTHORA SKLIRON CHROMA VASIS

Bad or Misleading Names

01A2;LATIN CAPITAL LETTER GHA
01A3;LATIN SMALL LETTER GHA
0285;LATIN SMALL LETTER REVERSED FISHHOOK R WITH RETROFLEX HOOK
0598;HEBREW ACCENT TSINNORIT
05AE;HEBREW ACCENT ZARQA (tsinor)
0670;ARABIC ALEF ABOVE
06C0;ARABIC LIGATURE HEH WITH YEH ABOVE
06C2;ARABIC LIGATURE HEH GOAL WITH HAMZA ABOVE
06D3;ARABIC LIGATURE YEH BARREE WITH HAMZA ABOVE
0B83;TAMIL AAYTHAM
0CDE;KANNADA LETTER LLLA
156F;CANADIAN SYLLABICS ASTERISK
2118;WEIERSTRASS ELLIPTIC FUNCTION
262B;SYMBOL OF IRAN
FEFF;BYTE ORDER MARK

The following are two sets of systematic usages that are not ideal, but since the usage is widespread and consistent in the character names, they are not as well suited for aliases

CARON -> HACEK

010C;LATIN CAPITAL LETTER C WITH HACEK
010D;LATIN SMALL LETTER C WITH HACEK
010E;LATIN CAPITAL LETTER D WITH HACEK
010F;LATIN SMALL LETTER D WITH HACEK
011A;LATIN CAPITAL LETTER E WITH HACEK
011B;LATIN SMALL LETTER E WITH HACEK
013D;LATIN CAPITAL LETTER L WITH HACEK
013E;LATIN SMALL LETTER L WITH HACEK
0147;LATIN CAPITAL LETTER N WITH HACEK
0148;LATIN SMALL LETTER N WITH HACEK
0158;LATIN CAPITAL LETTER R WITH HACEK
0159;LATIN SMALL LETTER R WITH HACEK
0160;LATIN CAPITAL LETTER S WITH HACEK
0161;LATIN SMALL LETTER S WITH HACEK
0164;LATIN CAPITAL LETTER T WITH HACEK
0165;LATIN SMALL LETTER T WITH HACEK
017D;LATIN CAPITAL LETTER Z WITH HACEK
017E;LATIN SMALL LETTER Z WITH HACEK
01C4;LATIN CAPITAL LETTER DZ WITH HACEK
01C5;LATIN CAPITAL LETTER D WITH SMALL LETTER Z WITH HACEK
01C6;LATIN SMALL LETTER DZ WITH HACEK
01CD;LATIN CAPITAL LETTER A WITH HACEK
01CE;LATIN SMALL LETTER A WITH HACEK
01CF;LATIN CAPITAL LETTER I WITH HACEK
01D0;LATIN SMALL LETTER I WITH HACEK
01D1;LATIN CAPITAL LETTER O WITH HACEK
01D2;LATIN SMALL LETTER O WITH HACEK
01D3;LATIN CAPITAL LETTER U WITH HACEK
01D4;LATIN SMALL LETTER U WITH HACEK
01D9;LATIN CAPITAL LETTER U WITH DIAERESIS AND HACEK
01DA;LATIN SMALL LETTER U WITH DIAERESIS AND HACEK
01E6;LATIN CAPITAL LETTER G WITH HACEK
01E7;LATIN SMALL LETTER G WITH HACEK
01E8;LATIN CAPITAL LETTER K WITH HACEK
01E9;LATIN SMALL LETTER K WITH HACEK
01EE;LATIN CAPITAL LETTER EZH WITH HACEK
01EF;LATIN SMALL LETTER EZH WITH HACEK

01F0;LATIN SMALL LETTER J WITH HACEK
01FC;LATIN CAPITAL LIGATURE AE WITH ACUTE
01FD;LATIN SMALL LIGATURE AE WITH ACUTE
021E;LATIN CAPITAL LETTER H WITH HACEK
021F;LATIN SMALL LETTER H WITH HACEK
02C7;MODIFIER LETTER HACEK
030C;COMBINING HACEK
032C;COMBINING HACEK BELOW
1E66;LATIN CAPITAL LETTER S WITH HACEK AND DOT ABOVE
1E67;LATIN SMALL LETTER S WITH HACEK AND DOT ABOVE

OPEN E -> EPSILON

0190;LATIN CAPITAL LETTER EPSILON
025B;LATIN SMALL LETTER EPSILON
025C;LATIN SMALL LETTER REVERSED EPSILON
025D;LATIN SMALL LETTER REVERSED EPSILON WITH HOOK
025E;LATIN SMALL LETTER CLOSED REVERSED EPSILON
029A;LATIN SMALL LETTER CLOSED EPSILON
1D08;LATIN SMALL LETTER TURNED EPSILON
1D4B;MODIFIER LETTER SMALL EPSILON
1D4C;MODIFIER LETTER SMALL TURNED EPSILON
1D93;LATIN SMALL LETTER EPSILON WITH RETROFLEX HOOK
1D94;LATIN SMALL LETTER REVERSED EPSILON WITH RETROFLEX HOOK
1D9F;MODIFIER LETTER SMALL REVERSED EPSILON