Re: Relations between certain Han properties

From: Mark Davis Date: 2015-10-18

Draft: https://goo.gl/RuLuEA

The following is a proposal for documenting and ensuring two invariants between properties.

1. [:Ideographic:] is a superset of [:Unified_Ideograph:].

I propose that we document that the containment as an invariant that programmers can depend on. (It would be a bizarre to have something be a [:Unified Ideograph:] yet not be [:Ideographic:].)

The 11 NFKC characters are: [〆〇 | -久]

- ✓ U+3006 IDEOGRAPHIC CLOSING MARK
- O U+3007 IDEOGRAPHIC NUMBER ZERO
- 9 HANGZHOU NUMERALs (all but 3)

1,002 CJK COMPATIBILITY IDEOGRAPHs (all)

3 HANGZHOU NUMERALs:

□ U+3038 HANGZHOU NUMERAL TEN toNFKC=+□ U+3039 HANGZHOU NUMERAL TWENTY toNFKC=+□ U+303A HANGZHOU NUMERAL THIRTY toNFKC=+

2. [:scx=Han:] is a superset of [:Ideographic:]

I propose documenting the containment as an invariant that programmers can depend on. All of the Ideographic are really *Han* Ideographic. *Remember that* [:scx=Han:] is the set of characters with Han as <u>at least one</u> of its SCX values: the characters in that set may have other scripts also.

BTW, [:scx=han:] contains 603 characters not in [: Ideographic:], namely [ロ-攵旡-ロー-龠、-〃々〈-】 = -、、。~※※※ と - 〇・「-^ ローロ(ー)-(至)(一)-優1月-12月0点-244平成-禁証1日-3旧、、。・・ローロ]. Most of these are compatibility (not NFKC), except for 210 outliers.

The 210 NFKC characters are: [ロ-攵旡-歺民-亀、- " 々〈-】 = -、 ーーーメンン と - 🛚 ・ 「 レローロ ヽ ヽローロ]

113 CJK RADICALs (all but 2).

- 36 CJK STROKES (all)
- 18 COUNTING ROD UNIT DIGITs (all)
- 2 IDEOGRAPHIC ANNOTATION characters (Kanbun)
- 41 misc punctuation/symbols

- 母 U+2E9F CJK RADICAL MOTHER
- □ U+2EF3 CJK RADICAL C-SIMPLIFIED TURTLE
- 214 KANGXI RADICALs (all)
- 14 IDEOGRAPHIC ANNOTATION characters (Kanbun, all but 2)
- 36 PARENTHESIZED IDEOGRAPHs (all)
- 49 CIRCLED IDEOGRAPHs (all)
- 12 IDEOGRAPHIC TELEGRAPH MONTHs (all)
- 25 IDEOGRAPHIC TELEGRAPH SYMBOL FOR HOURS (all)
- 31 IDEOGRAPHIC TELEGRAPH SYMBOL FOR DAYs (all)
- 4 SQUARE ERA NAMEs (all)
- 1 SQUARE CORPORATION (all)
- 5 HALFWIDTH punctuation variants

3. [:scx=han:] is almost a superset of [:name=/IDEOGRAPH/:]

The 12 NFKC characters are: [0-0]

12 IDEOGRAPHIC DESCRIPTION CHARACTERS (all)

Arguably these should be [:scx=Han:] and [:Ideographic:], since that's their fundamental function.

The 4 non-NFKC characters are: [\u3000 \ \ \ \ \ \ \ \]

All are [:scx=Common:] and [:Ideo=N:]

U+3000 IDEOGRAPHIC SPACE
U+FE11 PRESENTATION FORM FOR VERTICAL IDEOGRAPHIC COMMA
U+FE12 PRESENTATION FORM FOR VERTICAL IDEOGRAPHIC FULL STOP

U+FE51 SMALL IDEOGRAPHIC COMMA

Arguably the last 3 should be [:scx=Han:]; their corresponding NFKC characters are already [:scx=Han:]:

U+3001 IDEOGRAPHIC COMMA U+3002 IDEOGRAPHIC FULL STOP