Characters and CSS for Fonts

NTT Network Innovation Lab.

KAWABATA, Taichi

Prologue

In this presentation, I will explain 3 character/ fonts-related topics that may affect the standardization of CSS3.

- Character Display
  - Private Character and Font
  - IVS and Font Matching
- Character Search and Comparison
  - Normalization

Private Characters and Font Formats

Why Private Characters?

- Unicode Standard, Section 1.1

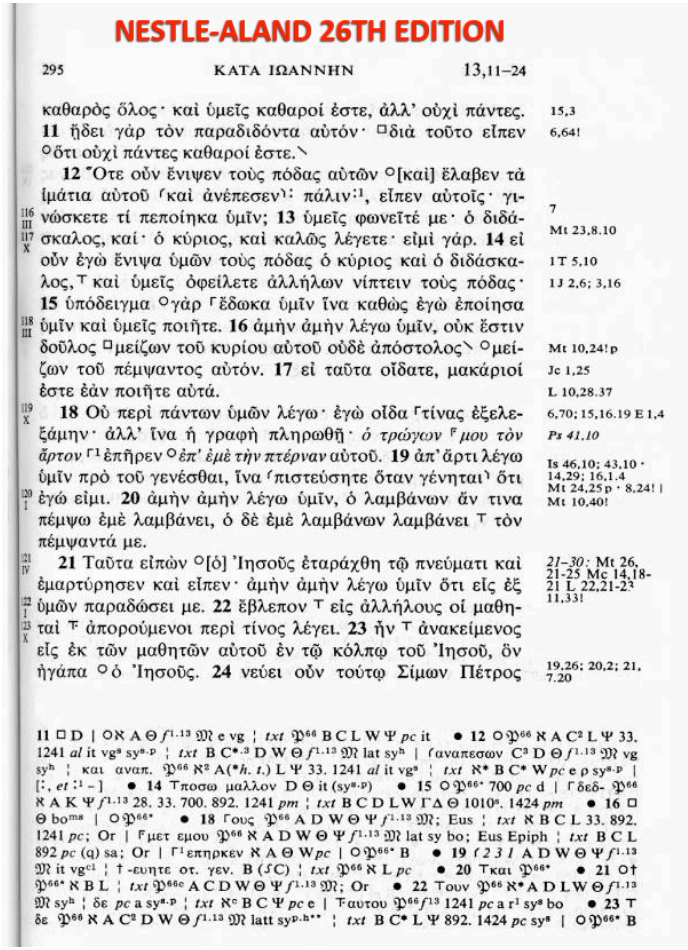
Note, however, that the Unicode Standard does not encode idiosyncratic, personal, novel, or private-use characters, nor does it encode logos or graphics. Graphologies unrelated to text, such as dance notations, are likewise outside the scope of the Unicode Standard. Font variants are explicitly not encoded. The Unicode Standard reserves 6,400 code points in the BMP for private use, which may be used to assign codes to characters not included in the repertoire of the Unicode Standard. Another 131,068 private-use code points are available outside the BMP, should 6,400 prove insufficient for particular applications.

What are Private Characters

- (Company) Logos and Graphics
- Unencoded characters
  - Emoticons (Emoji) and Special Symbols
- Some Kanji (CJK Ideograph) characters

Symbols and Emoticons

- In Books, special symbols are sometimes used to convey the complex or abstract idea in simpler mannar.
- Various Emoticons have become popular. Unicode begins to encode some of them, but not (yet) all of them.



NESTLE-ALAND 26TH EDITION

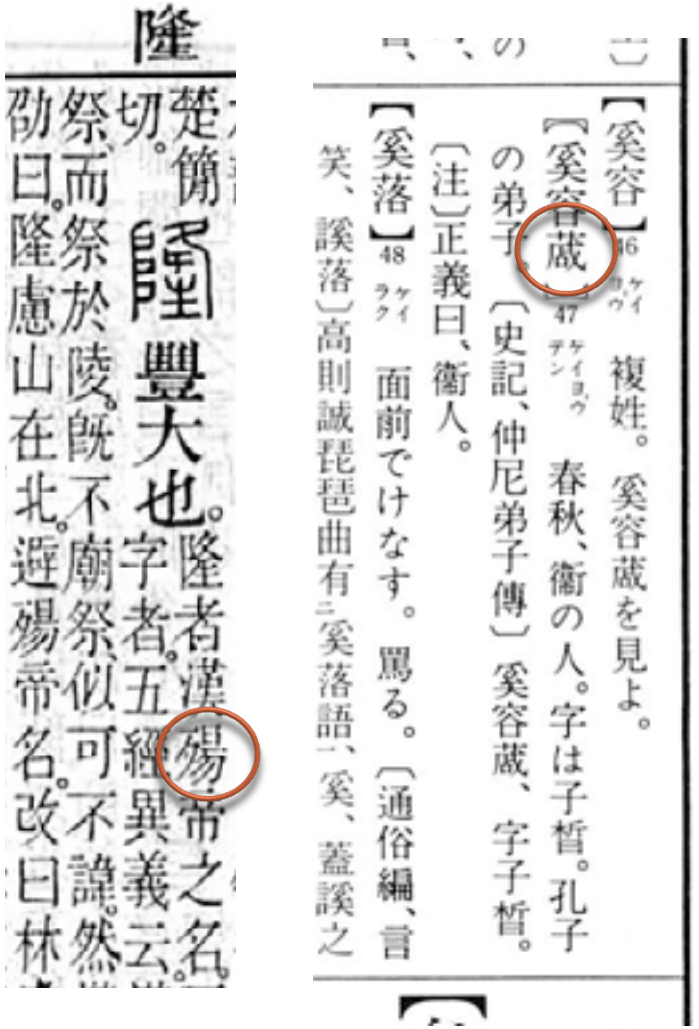
295 KATA IΩANNHN 13,11–24

καθαρὸς ὅλος· καὶ ὑμεῖς καθαροί ἐστε, ἀλλ' οὐχὶ πάντες. 15,3
11 ᾔδει γὰρ τὸν παραδιδόντα αὐτόν· □διὰ τοῦτο εἶπεν 6,64!
ὅτι οὐχὶ πάντες καθαροί ἐστε.
12 Ὅτε οὖν ἔνιψεν τοὺς πόδας αὐτῶν [καὶ] ἔλαβεν τὰ
ἱμάτια αὐτοῦ ⌜καὶ ἀνέπεσεν⌝· πάλιν· εἶπεν αὐτοῖς· γι- 7
116 III 117 X νώσκετε τί πεποίηκα ὑμῖν; **13** ὑμεῖς φωνεῖτέ με· ὁ διδά- Mt 23,8.10
σκαλος, καί· ὁ κύριος, καὶ καλῶς λέγετε· εἰμὶ γάρ. **14** εἰ
οὖν ἐγὼ ἔνιψα ὑμῶν τοὺς πόδας ὁ κύριος καὶ ὁ διδάσκα- 1T 5,10
λος, καὶ ὑμεῖς ὀφείλετε ἀλλήλων νίπτειν τοὺς πόδας· 1J 2,6; 3,16
15 ὑπόδειγμα γὰρ ⌜ἔδωκα ὑμῖν ἵνα καθὼς ἐγὼ ἐποίησα
118 III ὑμῖν καὶ ὑμεῖς ποιῆτε. **16** ἀμὴν ἀμὴν λέγω ὑμῖν, οὐκ ἔστιν Mt 10,24!p
δοῦλος □μείζων τοῦ κυρίου αὐτοῦ οὐδὲ ἀπόστολος⌝ μείζων Jc 1,25
τοῦ πέμψαντος αὐτόν. **17** εἰ ταῦτα οἴδατε, μακάριοί L 10,28.37
ἐστε ἐὰν ποιῆτε αὐτά.
119 X **18** Οὐ περὶ πάντων ὑμῶν λέγω· ἐγὼ οἶδα ⌜τίνας ἐξελε- 6,70; 15,16.19 E 1,4
ξάμην· ἀλλ' ἵνα ἡ γραφὴ πληρωθῇ· *ὁ τρώγων ⌜μου τὸν* *Ps 41,10*
ἄρτον ⌝ἐπῆρεν ἐπ' ἐμὲ τὴν πτέρναν αὐτοῦ. **19** ἀπ' ἄρτι λέγω Is 46,10; 43,10 · 14,29; 16,1.4
ὑμῖν πρὸ τοῦ γενέσθαι, ἵνα ⌜πιστεύσητε ὅταν γένηται⌝ ὅτι Mt 24,25 p · 8,24! |
120 I ἐγώ εἰμι. **20** ἀμὴν ἀμὴν λέγω ὑμῖν, ὁ λαμβάνων ἄν τινα Mt 10,40!
πέμψω ἐμὲ λαμβάνει, ὁ δὲ ἐμὲ λαμβάνων λαμβάνει τὸν
πέμψαντά με.
121 IV **21** Ταῦτα εἰπὼν [ὁ] Ἰησοῦς ἐταράχθη τῷ πνεύματι καὶ *21–30*: Mt 26, 21-25 Mc 14,18-
ἐμαρτύρησεν καὶ εἶπεν· ἀμὴν ἀμὴν λέγω ὑμῖν ὅτι εἷς ἐξ 21 L 22,21-23
122 I ὑμῶν παραδώσει με. **22** ἔβλεπον εἰς ἀλλήλους οἱ μαθη- 11,33!
123 X ταὶ περορούμενοι περὶ τίνος λέγει. **23** ἦν ἀνακείμενος
εἷς ἐκ τῶν μαθητῶν αὐτοῦ ἐν τῷ κόλπῳ τοῦ Ἰησοῦ, ὃν
ἠγάπα ὁ Ἰησοῦς. **24** νεύει οὖν τούτῳ Σίμων Πέτρος 19,26; 20,2; 21, 7.20

11 □D | ΟℵAΘ $f^{1.13}$ M e vg ¦ *txt* $\mathfrak{P}^{66}$ BCLWΨ *pc* it • **12** Ο$\mathfrak{P}^{66}$ ℵAC^{2}LΨ 33. 1241 *al* it vg^{s} sy$^{s.p}$ ¦ *txt* BC$^{*.3}$DWΘ $f^{1.13}$ M lat sy^{h} | ⌜αναπεσων C^{3}DΘ $f^{1.13}$ M vg sy^{h} ¦ και αναπ. $\mathfrak{P}^{66}$ ℵ^{2}A(**h. t.*)LΨ 33. 1241 *al* it vg^{s} ¦ *txt* ℵ* BC* W *pc* e ρ sy$^{s.p}$ | [:, *et* :^{1} –] • **14** Τποσω μαλλον DΘ it (sy$^{s.p}$) • **15** Ο$\mathfrak{P}^{66*}$ 700 *pc* d | ⌜δεδ- $\mathfrak{P}^{66}$ ℵAKΨ $f^{1.13}$ 28. 33. 700. 892. 1241 *pm* ¦ *txt* BCDLWΓΔΘ 1010^{s}. 1424 *pm* • **16** □ Θ bo^{ms} | Ο$\mathfrak{P}^{66*}$ • **18** ⌜ους $\mathfrak{P}^{66}$ ADWΘΨ $f^{1.13}$ M; Eus ¦ *txt* ℵBCL 33. 892. 1241 *pc*; Or | ⌜μετ εμου $\mathfrak{P}^{66}$ ℵADWΘΨ $f^{1.13}$ M lat sy bo; Eus Epiph ¦ *txt* BCL 892 *pc* (q) sa; Or | ⌜^{1}επηρκεν ℵAΘW *pc* | Ο$\mathfrak{P}^{66*}$ B • **19** *r 2 3 1* ADWΘΨ $f^{1.13}$ M it vg^{cl} ¦ †-ευητε οτ. γεν. B (*sC*) ¦ *txt* $\mathfrak{P}^{66}$ ℵL *pc* • **20** Τκαι $\mathfrak{P}^{66*}$ • **21** Ο† $\mathfrak{P}^{66*}$ ℵBL ¦ *txt* $\mathfrak{P}^{66c}$ ACDWΘΨ $f^{1.13}$ M; Or • **22** Τουν $\mathfrak{P}^{66}$ ℵ*ADLWΘ $f^{1.13}$ M sy^{h} ¦ δε *pc* a sy$^{s.p}$ ¦ *txt* ℵ^{c}BCΨ *pc* e | Ταυτου $\mathfrak{P}^{66}$ f^{13} 1241 *pc* a r^{1} sy^{s} bo • **23** Τ δε $\mathfrak{P}^{66}$ ℵAC^{2}DWΘ $f^{1.13}$ M latt sy$^{p.h**}$ ¦ *txt* BC*LΨ 892. 1424 *pc* sy^{s} | Ο$\mathfrak{P}^{66*}$ B

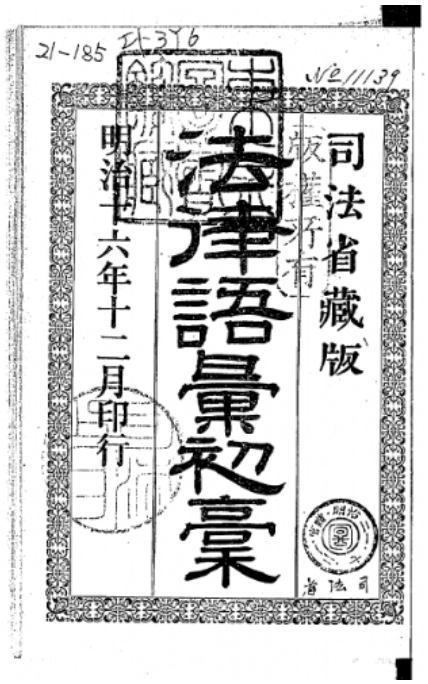
Kanji (CJK Ideographs)

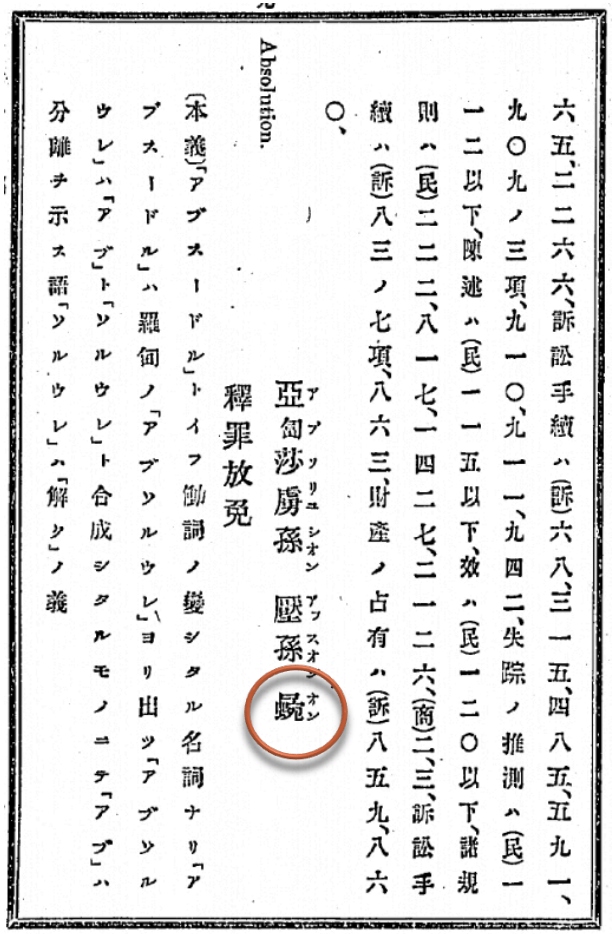
- 75,000+ ideographs are encoded under the Unification Rules.
- However, there are still more ideographs not yet encoded.
  - Misdescribed ideographs
  - Invented ideographs
  - (Very) Local ideographs
  - Historic, short-lived ideographs
  - Rare variants



Invented Ideographs

- Many ideographs are invented, but not widely used, in the past time.





In late 19th century, Meiji Government created hundreds of characters to describe new legal vocabraries, but they were never used.

Private Characters

- There are several ways to represent private characters

Methods	Prop	Search	Num	Editor	Notes
Assign Private Character to = (U+3013)	○	×	1	×	Not proper use
Assign Private Characer to ? (U+FFFD)	×	×	1	×	
Assign variants to encoded ideographs	○	○	1+	×	Only applicable to variants.
Use Private Use Area (PUA)	×	×	137,000+	◯	
Use inline image (<img>)	×	×	—	×	Hinting, etc. can not be applied.

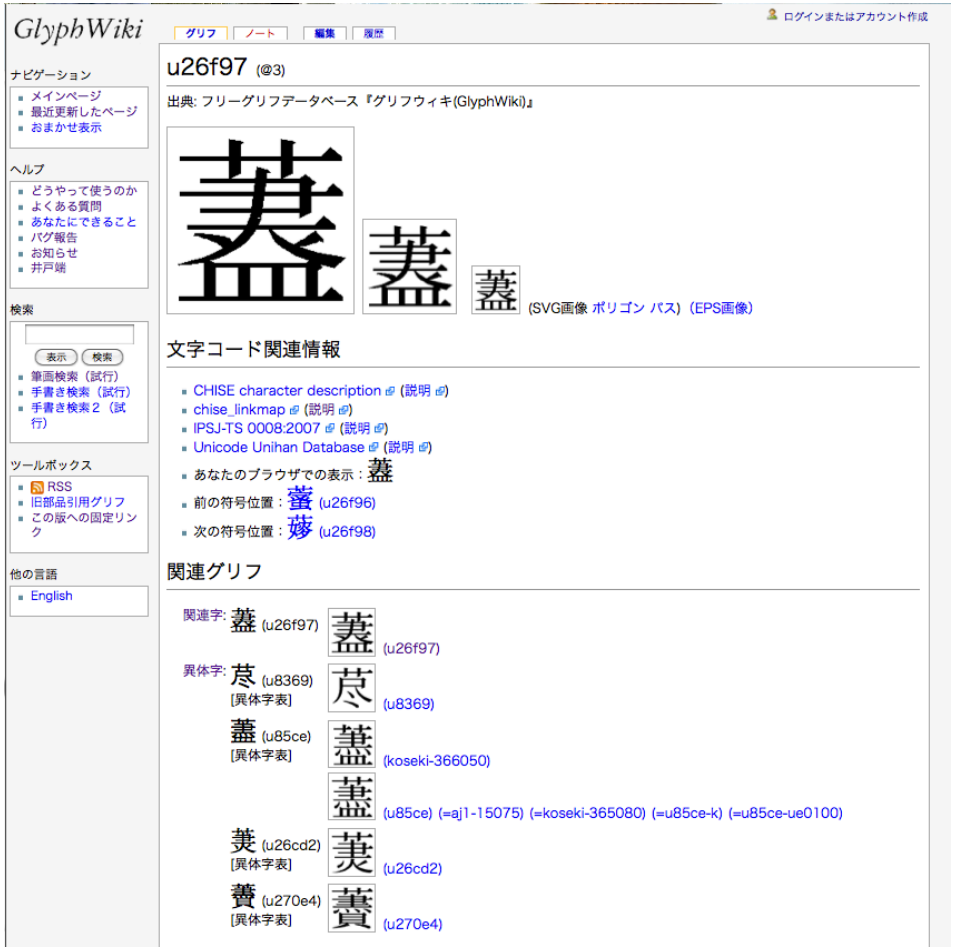
Font Formats

Font format	Pros	Cons
OpenType	Most popular. Various tools availabe.	Large Size. Piracy problem
WOFF	Small Size. Tailored for Web use.	You can't use it on OS.
SVG	gradation, color, animation Embeddable in HTML text. Can inherit CSS.	Not all browsers support it.

Emoticon Characters use colors.

GlyphWiki: Create/Share Private Characters

- http://glyphwiki.org/
- Everyone can edit/register his/her own characters.
- Fonts can be generated by creating independent wiki pages.
- Nearly 100,000 characters are registered.
- Hanazono-Mincho （花園明朝）is only one free font which covers all UCS/AJ1 ideographs.



Last UCS character registered to GlyphWiki

Private Chars and Text Orientation

- Character's behavior on different text orientation is determined by character's East Asian Width property.
- PUA characters do not have any property.
- One possible solution is to look at VORG/vert/vrt2 tables to determine character orientation.

When 'text-orientation' is 'vertical-right', set all characters upright (using vertical font settings if available) unless otherwise specified above.
In OpenType, vertical font settings are provided by the vhea, vmtx, and VORG tables, as well as the vert and vrt2 GSUB features. If any of these are present, the font is considered to have vertical font settings available. [CSS3 writing-mode]

IVS and Font Selection

Ideographic Variation Sequence (IVS)

- IVS enables to display Ideographic Variants by attaching “Variation Selector” after the “Base Character”.

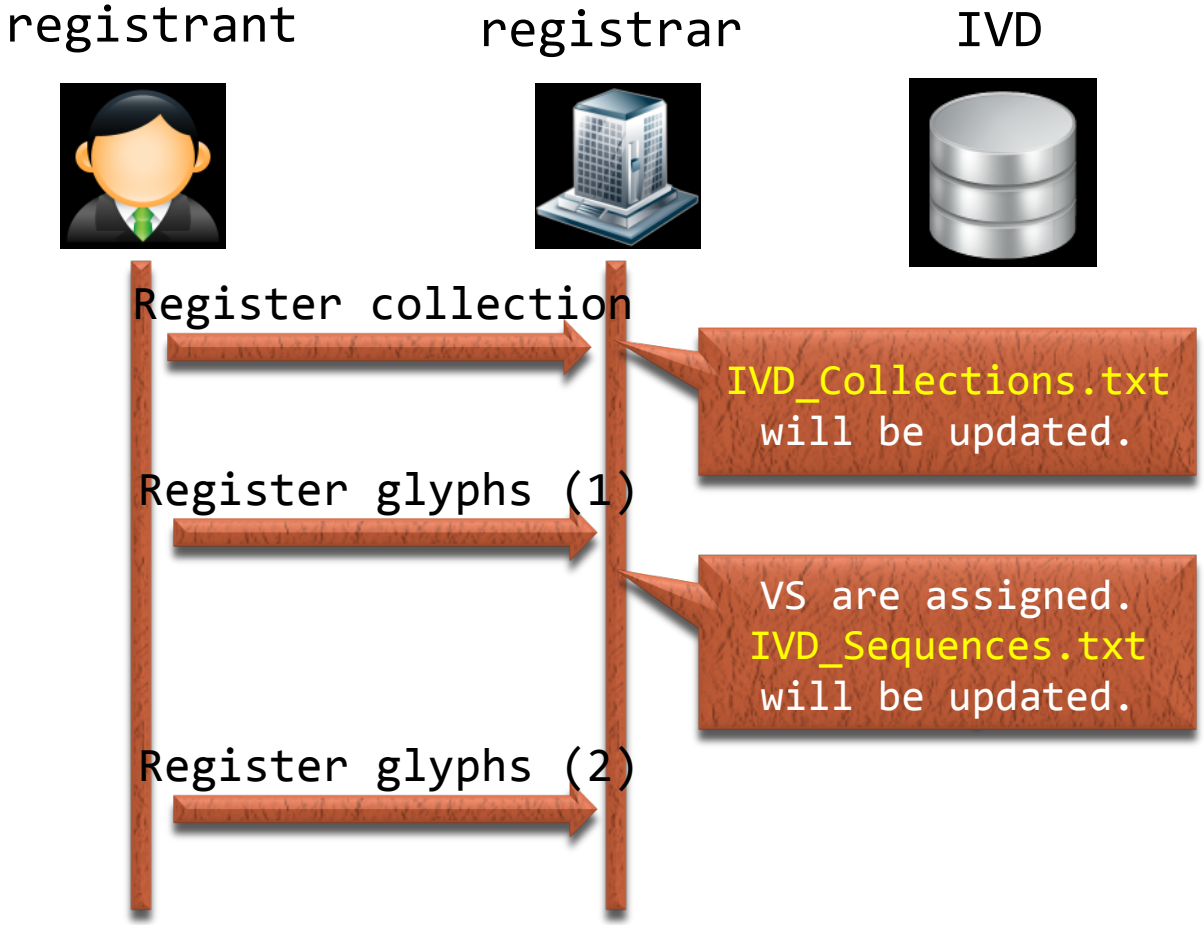
  e.g. “U+559D U+E0101” → 喝

  “U+559D U+E0100” → 喝

- Character code only specifies abstract character, but variation selector can specify concrete glyph for a character.

Ideographic Variation Database (IVD)

- UTS (Unicode Technical Standard) #37



IVD	Ideographic Variation Database
IVC	Ideographic Variation Collection
IVS	Ideographic Variation Selector
VS	Variation Selector

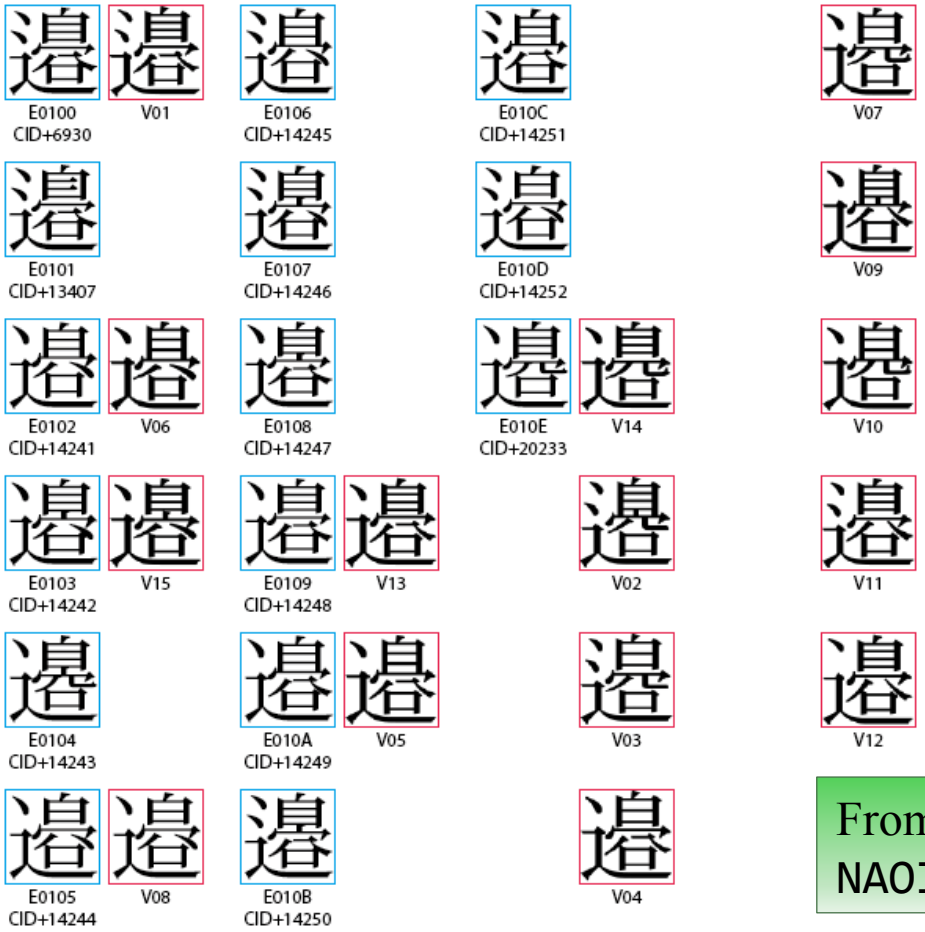
Ideographic Variation Collections

- Currently, Two collections are registered.

Collection Name	Purpose of Collection	Implemented Fonts available
Adobe-Japan1	To use with Japanese Desktop Publishing Systems.	Kozuka Mincho (小塚明朝)/Gothic(小塚ゴシック), etc
Hanyo-Denshi （汎用電子）	To use with Administration Systems of Japanese Government	IPAmjm Mincho

Glyph Correspondence

- 「邉」(AJ1: 15 glyphs, Hanyo-Denshi : 15 glyphs)



Not all variants are common between two collections

From http://d.hatena.ne.jp/NAOI/20100406/1270550459

Why IVS?

- Two main usage
  - To show “archaic style”
  - To correctly display human or proper names.

(Demonstrations)

CSS3 Font Matching Algorithm

- CSS font-family designation:

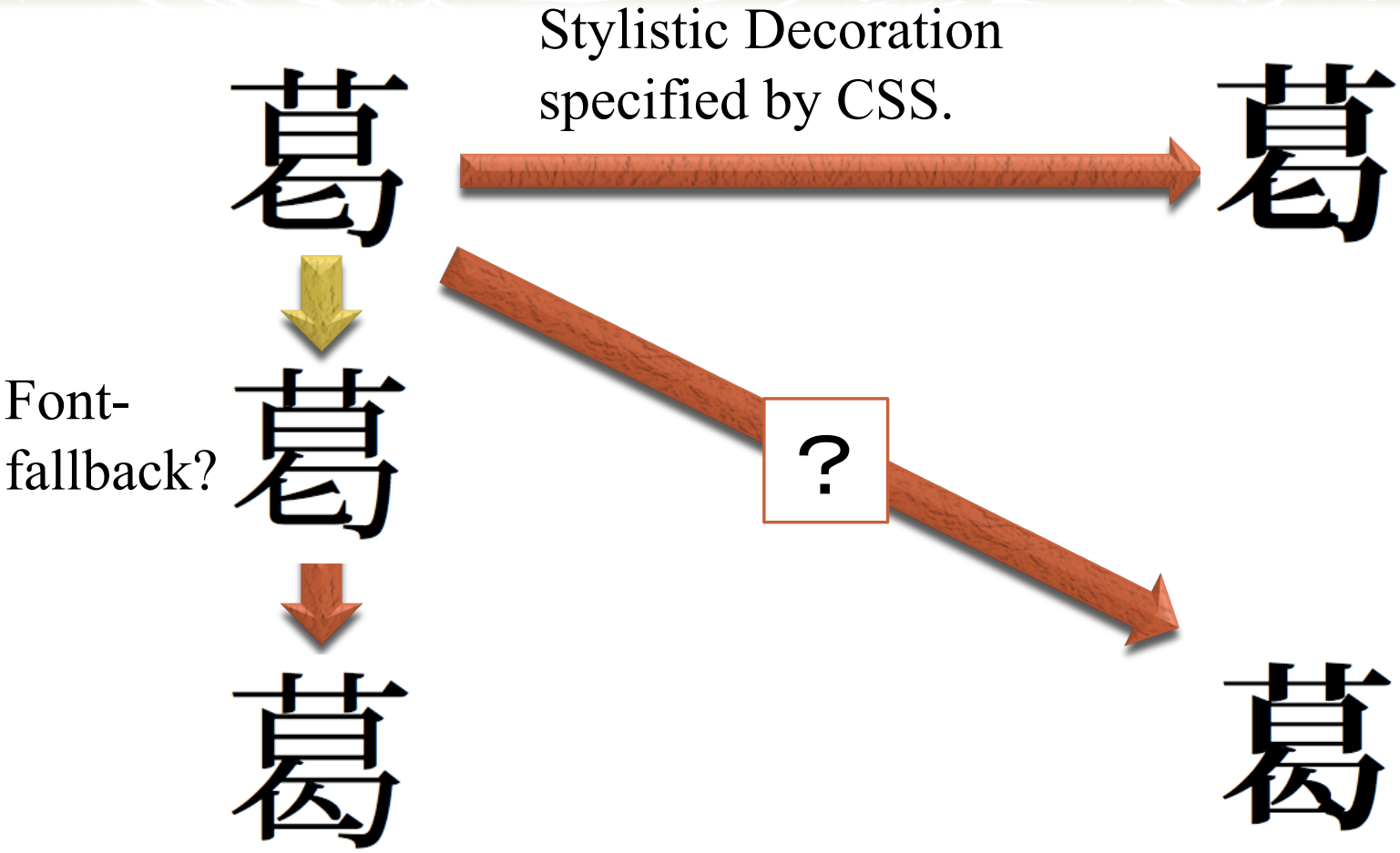
  `font-family:` *font-A, font-B, font-C;*

- Character (Grapheme Cluster) Sequence

  $C_1\ C_2\ C_3\ C_4\ C_5 \ldots$

- If fonts in *font-A* family contains all glyphs in C_1, then best font is selected by the following font spec order.
  - `font-stretch` (condensed/stretched)
  - `font-style` (italic/oblique)
  - `font-weight` (bold/light)

Two different text decorations



Variant specified by IVS.

Font Matching vs. IVS

- Consider CSS `font-family` designation:

  `font-family:` *font-A*, *font-B*, *font-C*;

  where *font-A*: supports only base characers.

  font-B: supports IVC_X

  font-C: supports IVC_Y

- Consider Character (IVS) Sequence:

  $C_1\ IVS_X\ (\in IVC_X)\ IVS_Y\ (\in IVC_Y)\ C_2\ \ldots$

- In which font family should IVS_X/IVS_Y be rendered?
  - Option A: $C_1\ BASE_X\ BASE_Y\ C_2$ (all by *font-A*)
  - Option B: $C_1\ IVSx\ IVSy\ C_2$ (by each IVS font.)

Which is better?

- Option A
  - Pro: Whole text has a consistent font-family.
  - Con: Multiple IVC fonts can not be supported.
- Option B
  - Pro: Each IVS will be rendered with their fonts.
  - Con: Text may be displayed in inconsistent font.

- Under option A, it is difficult for user to display each IVS(`font-family` must be specified in each IVS).
  - Font is specified by CSS font-family spec.
- Under option B, it is easy for user to display only base character (just remove VS characters) .
  - Variant is specified with VS character

Normalization

What is Normalization?

- Unicode has MANY ways to describe the same/similar character. Therefore, bitwise comparison do not guarantee correct result.
- e.g. U+1EB6 (NFC)
  U+1EA0 U+0306
  U+0041 U+0323 U+0306 (NFD)
  U+0102 U+0323
  U+0041 U+0306 U+0323
  are ALL the same character.

- If text is NORMALIZED, then text is guaranteed for comparison.
  - NFD … diacritical marks are all decomposed.
  - NFC … diacritical marks are composed in unique mannar
  - NFKC/NFKD … squared/circled chars are also decomposed.

Normalization Problem

- Implementation is cumbersome
  - NFC requires complex algorithm
- Singleton Decomposition
  - Different character will be folded to the same character. (It may go beyond specific subset)
  - Å（U+212B / JIS X 0208）→ Å（U+00C5 / ISO8859-1）
- Compatibility Ideographs
  - All compatibility ideographs will be transformed to corresponding unified ideographs.
    - Apple proposed to exclude Compatibility Characters from normalization, but it was not adopted. (10 years ago)

Compatibility Ideographs

- Ideographs are unified by unification rule specified in “ISO/IEC 10646 Annex S.”
  - Unifiable ideographs specified before 1992 are separately encoded into Unified Ideographs.

    (Source Code Separation Rule)

    e.g. 飲 (U＋98F2) vs. 飲 (U+98EE)
  - Unifiable ideographs specified after 1992 are assigned to Compatibility Ideographs.
- For example, Japanese compatibility ideographs include Name characters specified by the ordinance of the Ministry of Justice of Japan.

  e.g. 「社」vs.「社」/「者」vs.「者」

Normalization : When and Where?

- Early Unicode Normalization (2005)
  - **Character Model for the World Wide Web 1.0: Normalization**
  - Since then, it has become widely known that EUN has various problems.
    - http://www.w3.org/2011/05/04-i18n-minutes.html
- HTML Text should not be early normalized, but ID and Class values in HTML5 may/should be normalized.
  - XML specification Appendix J
- On web browser, "NFKC" normalization is more useful.
  - e.g. "ｼﾞｭｰｽ" can be searched by "ジュース"
  - e.g. "ﾘｯﾄﾙ" can be searched by "リットル"
- Old Kanji characters can not be searched by new characters. New normalization must be conceived.
  - e.g. 「小澤」 can not be searched by 「小沢」

Thanks for Listening!