CSSにおける文字とフォント

NTT未来ねっと研究所
川幡　太一

最初に

文字について、CSSの標準化に影響しそうな３トピック（外字とフォント、IVSとフォント選択、正規化）について、現在の状況と課題を、例を交えて説明する。

- 文字の表示
  - 外字とフォント
  - IVSとフォント選択
- 文字の照合
  - 正規化

外字とフォント

外字はなぜ必要か

- Unicode標準 第１章１節

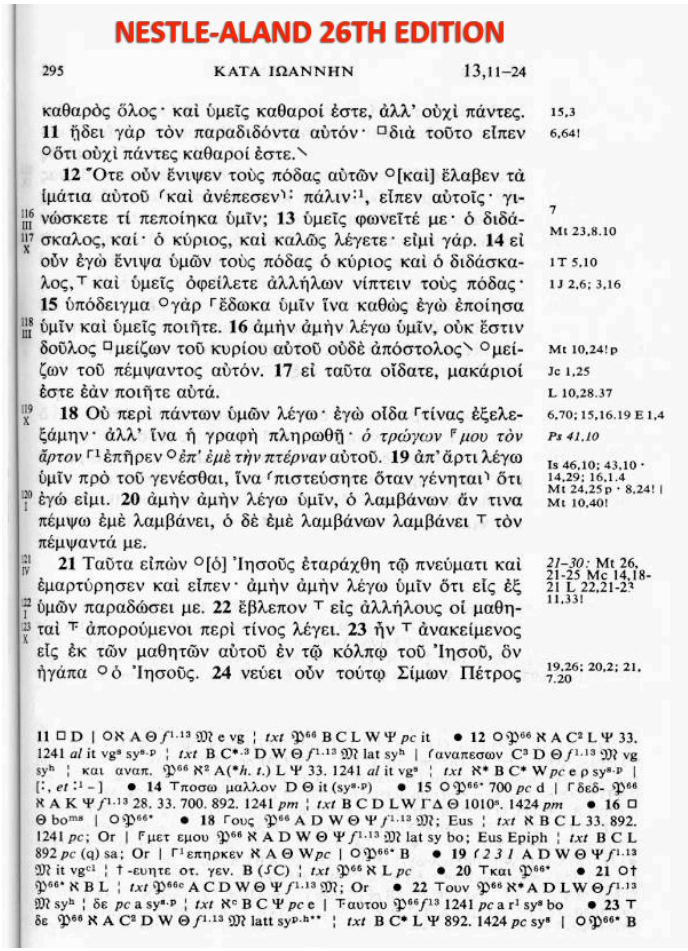
Note, however, that the Unicode Standard does not encode idiosyncratic, personal, novel, or private-use characters, nor does it encode logos or graphics. Graphologies unrelated to text, such as dance notations, are likewise outside the scope of the Unicode Standard. Font variants are explicitly not encoded. 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.

外字の種類

- 会社のロゴやシンボルマーク
- 未符号化文字
  - 絵文字・記号類
- 未符号化の漢字

記号類・絵文字

- 繰り返し現れる項目や、複雑な概念を抽象化・簡潔化して表記するため、独自文字を使用。
- 日本の携帯電話の絵文字


NESTLE-ALAND 26TH EDITION

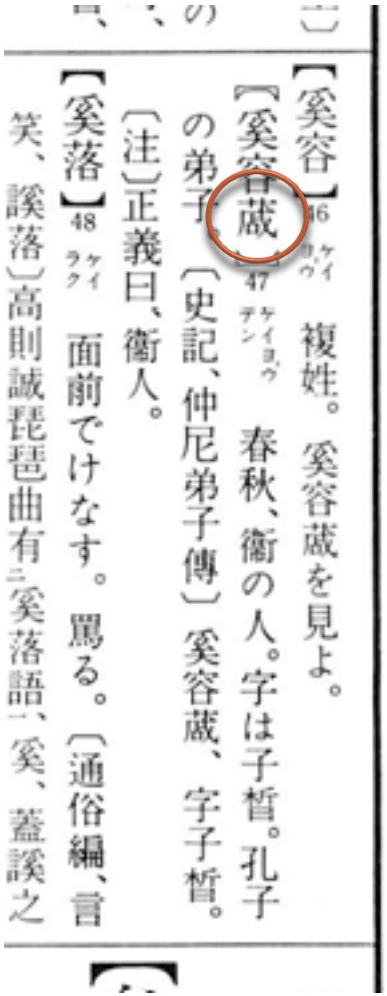
295 KATA IΩANNHN 13,11–24

καθαρὸς ὅλος· καὶ ὑμεῖς καθαροί ἐστε, ἀλλ' οὐχὶ πάντες. 15,3
11 ἤδει 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 ὁ ἐπ' ἐμὲ τὴν πτέρναν αὐτοῦ. 19 ἀπ' ἄρτι λέγω Is 46,10; 43,10 ·
ὑμῖν πρὸ τοῦ γενέσθαι, ἵνα ⌜πιστεύσητε ὅταν γένηται⌝ ὅτι 14,29; 16,1.4
120 I ἐγώ εἰμι. 20 ἀμὴν ἀμὴν λέγω ὑμῖν, ὁ λαμβάνων ἄν τινα Mt 24,25 p · 8,24! |
πέμψω ἐμὲ λαμβάνει, ὁ δὲ ἐμὲ λαμβάνων λαμβάνει τὸν Mt 10,40!
πέμψαντά με.
121 IV 21 Ταῦτα εἰπὼν ὁ [ὁ] Ἰησοῦς ἐταράχθη τῷ πνεύματι καὶ *21–30*: Mt 26,
ἐμαρτύρησεν καὶ εἶπεν· ἀμὴν ἀμὴν λέγω ὑμῖν ὅτι εἷς ἐξ 21-25 Mc 14,18-
122 I ὑμῶν παραδώσει με. 22 ἔβλεπον εἰς ἀλλήλους οἱ μαθη- 21 L 22,21-23
123 X ταὶ ἀπορούμενοι περὶ τίνος λέγει. 23 ἦν ἀνακείμενος 11,33!
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* 𝔓66 B C L W Ψ *pc* it • 12 Ο𝔓66 ℵ A C^{2} L Ψ 33. 1241 *al* it vg^{s} sy$^{s.p}$ ¦ *txt* B C$^{*.3}$ D W Θ $f^{1.13}$ 𝔐 lat sy^{h} | ⌜αναπεσων C^{3} D Θ $f^{1.13}$ 𝔐 vg sy^{h} ¦ και αναπ. 𝔓66 ℵ^{2} A(**h. t.*) L Ψ 33. 1241 *al* it vg^{s} ¦ *txt* ℵ* B C* W *pc* e ρ sy$^{s.p}$ | [:, *et* :1 –] • 14 Τποσω μαλλον D Θ it (sy$^{s.p}$) • 15 Ο𝔓66* 700 *pc* d | ⌜δεδ- 𝔓66 ℵ A K Ψ $f^{1.13}$ 28. 33. 700. 892. 1241 *pm* ¦ *txt* B C D L W Γ Δ Θ 1010^{s}. 1424 *pm* • 16 □ Θ bo^{ms} | Ο𝔓66* • 18 ⌜ους 𝔓66 A D W Θ Ψ $f^{1.13}$ 𝔐; Eus ¦ *txt* ℵ B C L 33. 892. 1241 *pc*; Or | ⌜μετ εμου 𝔓66 ℵ A D W Θ Ψ $f^{1.13}$ 𝔐 lat sy bo; Eus Epiph ¦ *txt* B C L 892 *pc* (q) sa; Or | ⌜¹επηρκεν ℵ A Θ W *pc* | Ο𝔓66* B • 19 *f 2 3 1* A D W Θ Ψ $f^{1.13}$ 𝔐 it vg^{cl} ¦ †-ευητε οτ. γεν. B (*sC*) ¦ *txt* 𝔓66 ℵ L *pc* • 20 Τκαι 𝔓66* • 21 Ο† 𝔓66* ℵ B L ¦ *txt* 𝔓66c A C D W Θ Ψ $f^{1.13}$ 𝔐; Or • 22 Τουν 𝔓66 ℵ* A D L W Θ $f^{1.13}$ 𝔐 sy^{h} ¦ δε *pc* a sy$^{s.p}$ ¦ *txt* ℵc B C Ψ *pc* e | Ταυτου 𝔓66 f^{13} 1241 *pc* a r^{1} sy^{s} bo • 23 Τ δε 𝔓66 ℵ A C^{2} D W Θ $f^{1.13}$ 𝔐 latt sy$^{p.h**}$ ¦ *txt* B C* L Ψ 892. 1424 *pc* sy^{s} | Ο𝔓66* B

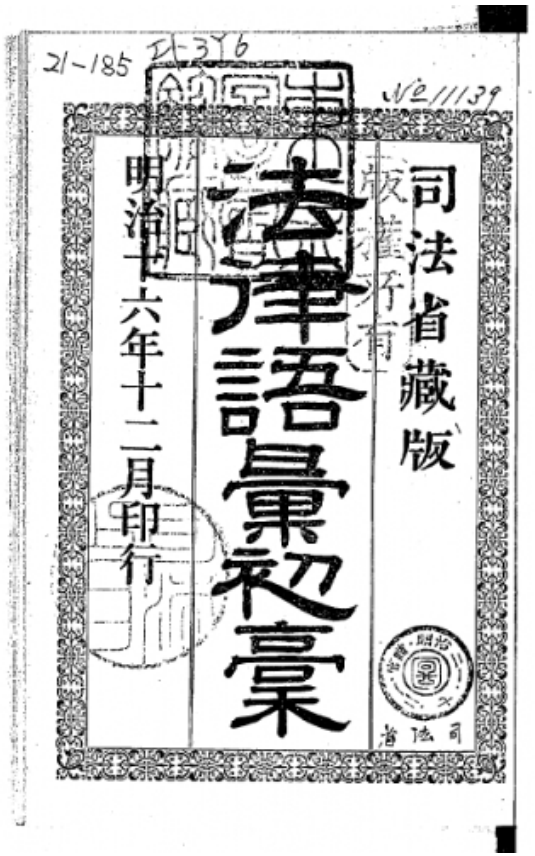
漢字

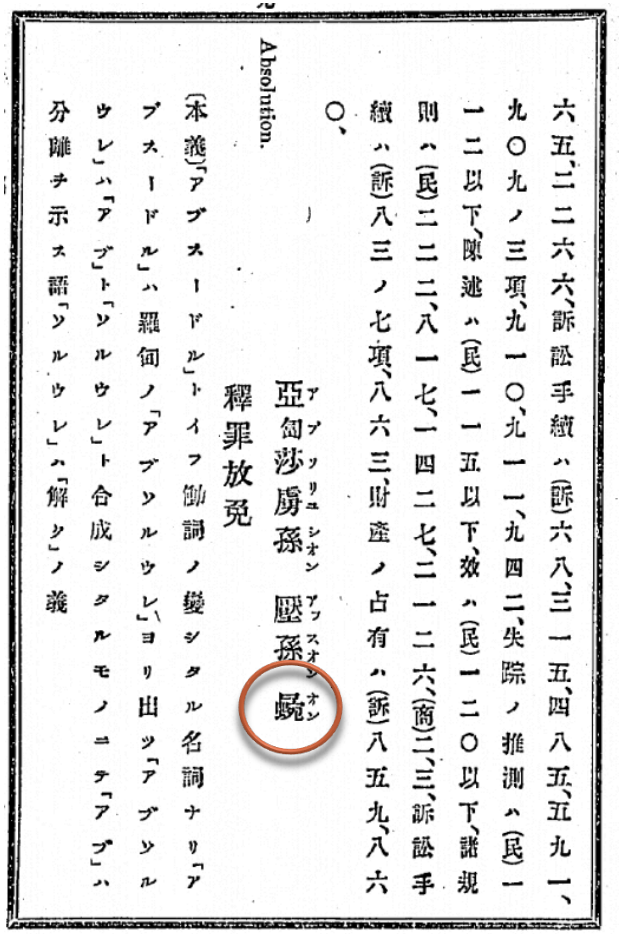
- ７万５千字以上が「統合基準」に基づいて符号化
- まだ多くの漢字が未符号化。
  - 誤字
  - 人工文字
  - 特定の地域でのみ使われる文字
  - 歴史的な文字
  - 珍しい異体字



新造漢字

- 漢字はしばしば、新規に作字される。





「法律語彙初稿」には、明治政府が法律語彙用に試作した数百の新造漢字が掲載されている。

外字

- CSS/HTML で外字を表現する方法はいくつかある。

方法	属性	検索	数	エディタ	メモ
〓 (U+3013)に外字を割り当てる	○	×	1	×	正しい使い方ではない。エディタで文字が見えない。
□ (U+FFFD)に外字を割り当てる	×	×	1	×	縦書きかが判別しない。
符号化された漢字に異体字を割り当てる	○	○	1	×	異体字にしか適用できない。
私用文字領域 (PUA) を使う	×	×	137,000+	◯	文字の属性が不明
インライン画像 (<img>)を使う	○	×	ー	×	ヒント情報が使えない。

外字フォント

フォント	メリット	デメリット
OpenType	ツールが充実	サイズが大きくなりがち ライセンス上、Webに埋め込むのは困難な場合がある。
WOFF	サイズが小さい。 Webで利用しやすい	OSではそのままでは使えない。
SVG	グラデーション・色・アニメが可能 HTMLに埋め込み可能。 CSSを継承することが可能。	ヒントや、複雑なリガチャ・カーニングなどに未対応 EPUBで必ずしもサポートされない。

GlyphWiki: 外字の共有

- 誰でもグリフを作成・編集して登録可能
- フォントは新規Wikiページを作成することで生成。
- 100,000 以上のグリフ（主に漢字）が登録済み
- 花園明朝は世界で唯一、全UCS漢字とAdobe-Japan1グリフをサポートしているフリーフォント

外字と縦書き

- 縦書きの時に、文字を倒すかは、文字のEAW属性などを使って決定する。
- 符号位置から、文字の性質が分からない場合、縦書き時に正立するか？

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.

IVSとフォント選択

漢字字形指示列（IVS）

- IVSは、漢字の基底文字の符号の直後に、異体字選択子（VS）を付加することで、異体字の表示を可能にする。

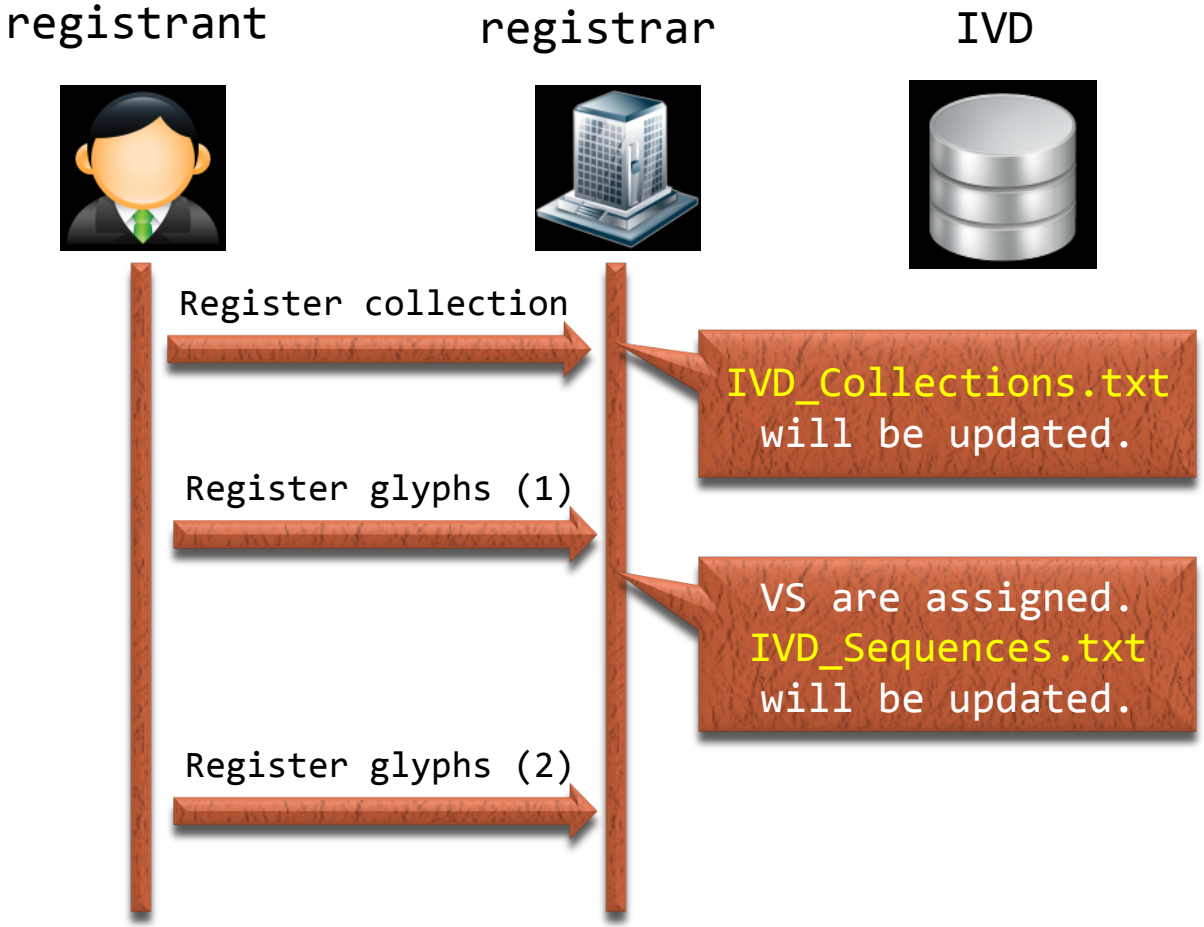
  e.g.　“U+559D U+E0101” → 喝

  “U+559D U+E0100” → 喝

- 文字コードは、抽象的な文字しか指示しないが、異体字選択子は直接、具体的なグリフを指示する。

漢字字形データベース（IVD）

- UTS (Unicode Technical Standard) #37



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

異体字コレクション（IVC）

- 現在、２つのコレクションが登録されている。

コレクション名	コレクションの目的	実装済のフォント
Adobe-Japan1	日本語のDTPでの使用	小塚明朝 小塚ゴシック
Hanyo-Denshi （汎用電子）	日本の行政システムでの使用	IPAmjm 明朝

各コレクションのグリフ対応例

- 「邉」(AJ1: 15 グリフ, 汎用電子: 15 グリフ)

http://d.hatena.ne.jp/NAOI/20100406/1270550459
より引用

なぜIVS?

- ２つの主な用途
  - テキストを「古風」に見せる（スタイル）
    - （従来のOpenTypeのjp78のような使い方）
  - 人名や固有名詞を適切に表示する。

（Demonstration）

CSS3 フォント選択アルゴリズム

- CSS `font-family` 指定

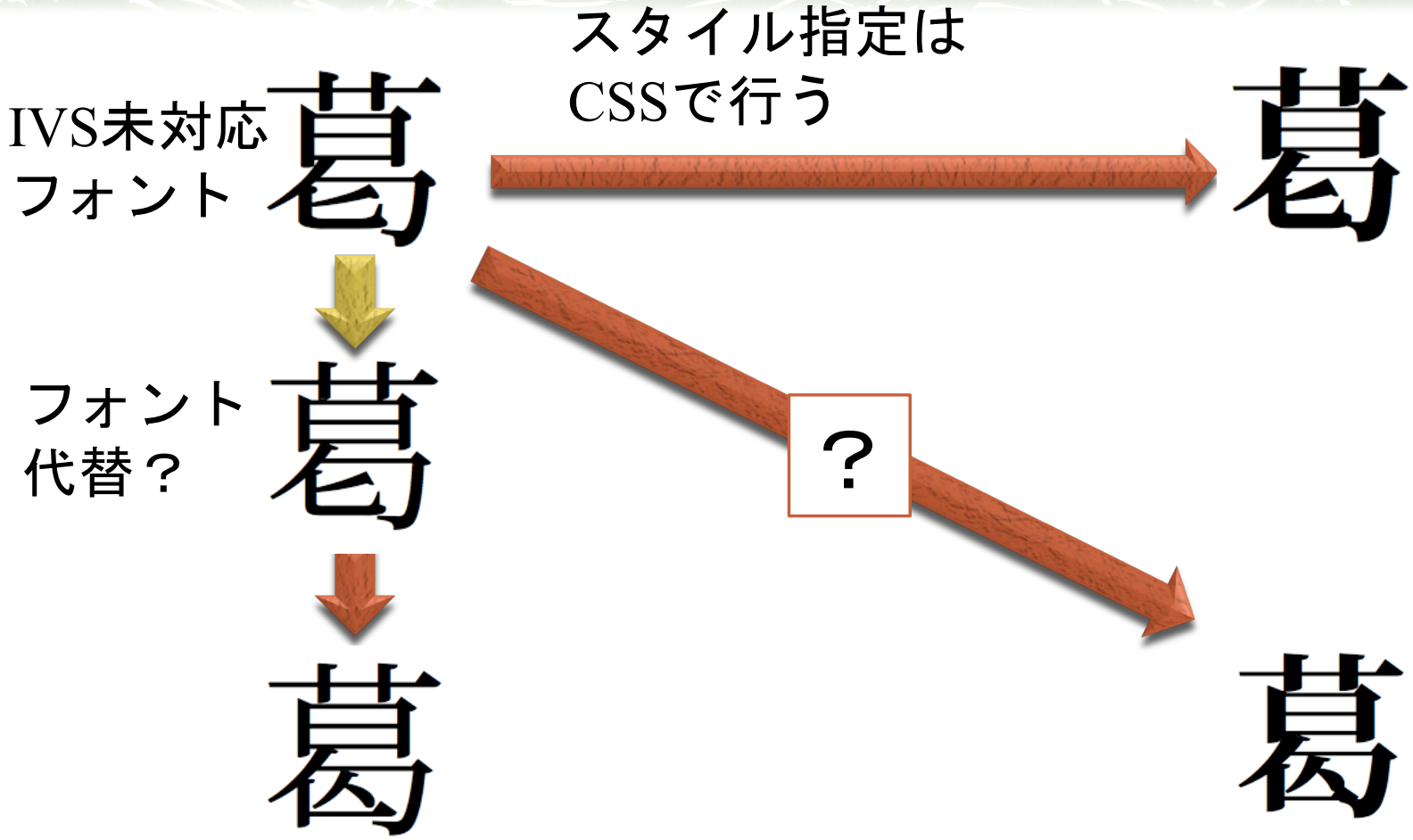
  `font-family:` *font-A, font-B, font-C;*

- 文字（書記素クラスタ）列

  $C_1\ C_2\ C_3\ C_4\ C_5 \cdots$

- *font-A* ファミリで指定されたフォント群が、 C_1に含まれる全グリフをサポートしている場合、以下の優先度でフォントが選択される。
  - `font-stretch` (condensed/stretched)
  - `font-style` (italic/oblique)
  - `font-weight` (bold/light)

２つの異なるテキスト装飾



異体字はIVSで指定する。

IVSとフォント選択

- CSS `font-family` 指定:

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

  font-A: 基底文字のみ搭載

  font-B: 異体字コレクションXをサポート

  font-C: 異体字コレクションYをサポート

- 文字・書記素（IVS） 列:

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

- どのフォントファミリで、IVS_X/IVS_Y は表示されるべきか？
  - A案: $C_1\ BASE_X\ BASE_Y\ C_2$ (all by *font-A*)
  - B案: C_1 IVSx IVSy C_2 (by each IVS font.)

IVSはどちらで表示するべきか？

- A案
  - 利点：テキスト全体でフォントの一貫性が保たれる.
  - 欠点：複数のIVCフォントを混在できない。
- B案
  - 利点：各IVSはそれぞれに対応したフォントで表示可能。
  - 欠点：テキスト全体でフォントの一貫性が欠けるかも。

- A案では、ユーザが複数のIVCを混在させるのは困難。（IVS毎にfont-familyを指定する必要がある）。
  - フォントを指定するのはCSS。
- B案では、ユーザが基底文字のみを表示するのは簡単。（単にVSを削除すれば良い。）
  - 異体字を選択するVSは文字符号。

正規化について

正規化とは何か？

- Unicodeは、同じ文字を表現するのに多数の方法がある。そのため、文字列をビット単位で比較しても、正しい比較はできない。
- 例：U+1EB6 (NFC)
  U+1EA0 U+0306
  U+0041 U+0323 U+0306 (NFD)
  U+0102 U+0323
  U+0041 U+0306 U+0323
  はすべて同じ文字。

- 事前に正規化すれば、ビット比較が正しいことが保証。
  - NFD … 補助記号類はすべて分解される。（かなの濁音等）
  - NFC … 補助記号類は同一性を保証する形で合成される。
  - NFKD/NFKC … 丸付き文字（①など）や組文字（㍑など）も分解（＋再合成）される。

なぜ正規化は必要か？

- 文字列を正しく比較できる。
  - 識別子で使う文字は、正規化しないとマッチングが取れない。
- HTML/CSSでは特に、id値、class名などは正規化が必要。
- 正規化は万能か？

正規化の課題

- 実装（特にNFC）が複雑
- 単一分解 (Singleton Decomposition)
  - 異なる文字が、同じ文字に畳み込まれる場合がある。(その結果、指定された部分集合外に変換される場合がある。)
    - 例：Å（U+212B / JIS X 0208）→ Å（U+00C5 / ISO8859-1）
- 互換漢字
  - すべての互換漢字は、正規化によって対応する統合漢字に変換される。
  - この挙動には問題が多く、Apple等が除外を求めたが受け入れられず。

互換漢字

- UCSの漢字は、ISO/IEC 10646 追補S で記述された規則にしたがって統合される。
  - 1992年以前に規格化された統合可能な漢字は、統合漢字で分離して符号化（原規格分離規則）
    - 例：「飲」と「飲」は本来統合されて１符号になるはずが、JISの都合により分離し符号化。
  - 1992年以降の統合可能な漢字は、互換漢字で符号化。
- 互換漢字の例
  - 法務省が省令で定めた人名用漢字表の漢字（JIS X 0213）
  - 例：「社」と「社」や、「者」と「者」

正規化は、いつ、どこで行うか？

- 事前正規化…Early Unicode Normalization (2005)
  - **Character Model for the World Wide Web 1.0: Normalization**
  - HTMLテキストを必ず事前に正規化する提案だが、現在は課題が多くの人に認識されつつある。
    - http://www.w3.org/2011/05/04-i18n-minutes.html
- テキストの事前正規化は不適切だが、HTMLのID値とClass値 の正規化は適当かもしれない。
  - XML specification Appendix J
- ブラウザ側で “NFKC” 正規化が便利。
  - 例：「ｼﾞｭｰｽ」が「ジュース」で検索できる。
    「ﾘｯﾙ」が「リットル」で検索できる。
- 漢字の旧字は未だに新字で検索不可。何か仕組が必要
  - 例：「小澤」は「小沢」で検索できない。

Thanks for Listening!