

区別することの必要性を明確化することの 必要性ということについて

小林 龍生
有限会社 スコレックス

筆者は 1995 年以来国際符号化文字集合の標準化活動に関わってきた。この過程で、文字集合として符号化する文字の異同を判断する粒度の基準が、言語文化や使用目的によって異なることに起因する合意形成の困難さと対峙し続けてきた。この問題解消の一助とするため、VS(Variation Selector)というメカニズムの提案および規格開発に主体的に関わってきた。本稿では、これらの経験を踏まえ、標準規格策定および実用システム開発に当たっては、利用者側が利用意図を明確に意識した上で、文字異同の判断基準を明確にすることの重要性について述べる。

Importance of the clarification of the difference among similar character shapes

Tatsuo Kobayashi
Scholex Co., Ltd.

I have been involved in the development of ISO/IEC 10646, Universal Coded Character Set from 1995. In this work I have been bothered by the difference of granularity of glyph shape determination among different cultures and usages. Through this experience, I proposed and developed VS (Variation Selector) mechanism with colleagues. In this paper I discuss the importance of clarification of granularity of glyph determination by the user for their usage.

1. まえがき

情報通信機器を用いた文字による情報交換においては、通常、符号化文字集合、いわゆる文字コードが用いられる。しかし、ある符号化文字集合を用いた情報交換が、情報の送り手と受け手の間で、常に十全に行われるわけではない。さまざまなレベルでの文字化けや外字の必要性が端的に物語るように、符号化文字集合における情報交換に際しては、情報の送り手と受け手の間で、用いる文字集合についての事前の合意が不可欠である。この事前の合意を、ある社会的な集団の範囲で一般化したものこそが、符号化文字集合標準である。

しかし、言語文化の多様性、利用目的の多様性を勘案すると、一つの普遍的な文字集合で、すべての言語文化、利用目的に対応することがきわめて困難なこともまた言を俟たない。

本稿では、筆者の実体験を踏まえた上で、利用者側において利用目的を明確化することの重要性と、明確にされた利用目的を踏まえた上で、規格策定者側、実装技術側での解決策の提示の可能性について議論する。

2. Character Glyph Model

ISO/IEC JTC1(Joint Technical Committee)では、SC2 が Coded Character Set に係わる規格策定を担当しており、SC34 が、Glyph および Font に関

わる規格策定を担当している。他の分野同様、文字および文書周辺の国際規格策定に際しても、さまざまなレベルでの組織間の軋轢がある。SC2 と SC34 では、それぞれの担当範囲を明確化するために、共同で一つの Technical Report の開発を行った。それが、いわゆる Character Glyph Model と呼ばれる、ISO/IEC TR 15285 Information technology --- An operational model for characters and glyphs である。

以後の議論は、この Character Glyph Model に示されている定義を前提として行う。

character: A member of a set of elements used for the organization, control, or representation of data. (ISO/IEC 10646-1:1993)

glyph: A recognizable abstract graphic symbol which is independent of any specific design. (ISO/IEC 9541-1: 1991)

ここで、ISO/IEC 10646[1]（本項では、UCS という略称を用いる場合もある）の Character（ISO/IEC 10646 を翻訳した JIS X 0221 では《文字》と訳している）の定義を見ておこう。JIS X 0221 の翻訳では、“データの構成、制御又は表現に用いる要素の集合の構成単位”となってる。この定義には、どこにも一般的な意味での文字に相当する概念が含まれていない。この定義を筆者なりにパラフレーズすると、UCS における

《文字》とは、“ある集合の中で排他的に一意的な固有のビット列もしくはそのビット列に付けられた名前”ということになる。

一方、Glyph の方は、比較的簡単で、特別なデザインに拘束されない抽象的な視覚的図像とでも理解しておけばよい。巷間、Glyph が、微細なデザインの差異も含めた具体的な図形の視覚表現の意味で用いられる例も見られるが、本稿では、Glyph はあくまでも抽象的な図像の意味で用いる。

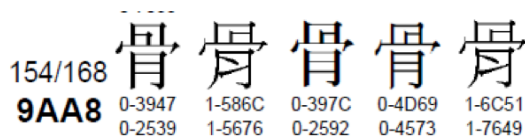
昭和 24 年 4 月 28 日に内閣告示第 1 号として公布された“当用漢字字体表”では、字形が（おそらくは手書きの）細い鉛筆書きのような文字で示されているが、このような表記方法が用いられた意図を事後的に解釈すると、ここではまさに“Glyph”が示されていると言えよう。

Character Glyph Model の観点からは、符号化文字集合とは、単一または複数の Glyph に対応する固有のビット列もしくは名前を定めたもの、と言うことが出来よう。

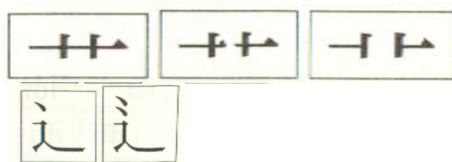
3. 包摂規準もしくは統合規則

符号化文字集合は、単一もしくは複数の Glyph に対応する固有名の集合である。日本語のための漢字集合である JIS X 0208 や JIS X 0213 では包摂規準として、ISO/IEC 10646 では“Unification Rule”（統合規則）^[2]として、一つの符号位置に対応させる Glyph の範囲を例示している。

UCS では、1990 年代半ばに吹き荒れたユニコード批判の矢面に立たされた《骨》の例が典型的なものである。下記の図は、ISO/IEC 10646 の統合漢字ブロックの一部である。左から中国、台湾、日本、韓国、ヴェトナムの順に、国内規格の例示字形が掲載されている。中国の欄を見ると、骨の字の上部が、台湾や日本の字形と異なっていることが分かる。^[3]



JIS X 0208 では、草冠や之繞などが挙げられる。下記の例は、JIS X 0208 の包摂規準を規定した部分の一部であるが、3 画の草冠と 2 種類の 4 画の草冠、1 点之繞と 2 点之繞が包摂されている。^[4]



この包摂規準と統合規則の相違が、複数の文字集合間の相互運用性に大きな障害となることもまた言を俟たない。

一例を挙げると、森鷗外の名前で有名な《鷗》と《鷗》がある。これらは、JIS X 0208 では包摂規準により包摂されていたものが、JIS X 0213 では包摂分離されてそれぞれが別の符号位置に対応する。この例では、例えば、JIS X 0208 の環境下である特定のフォントを用いて入力した“森鷗外”が、同じ JIS X 0208 環境下で別のフォントを用いてプリントアウトされた際に“森鷗外”と印字される場合や、JIS X 0213 の環境下で入力された“森鷗外”が JIS X 0208 の環境下では“森＝外”と出力される場合などが想定される。

当然のことながら、日本語の表記に限っても、包摂規準も相対的なものとならざるを得ず、利用目的によっては粒度が細かすぎるという批判も存在しうる。粒度を細かくすることの功罪については、後述する。

4. Glyph 異同の粒度

粒度という言葉は、現時点ではまだ日本語として一般化しているとは言いがたいかも知れない。ここでは、granularity に対応する言葉として用いる。

定義としては明確な Glyph だが、実際には、Glyph の異同についても、文化や用途によって、かなり大きな粒度の差が存在する。

後述する IVD には、現時点で、Adobe Japan1 というコレクションと Hanyo-Denshi というコレクションの二つが登録されている。Adobe Japan1 コレクションは、主として JIS 漢字符号の変遷と日本の印刷業界で伝統的に用いられてきた Glyph を反映したものとなっており、Hanyo-Denshi コレクションは、戸籍統一文字と住基ネット統一文字という二つの Glyph セットを整理統合した汎用電子情報交換環境整備プログラムの成果の一部が登録されている。この二つのコレクションを比較しただけでも、出版・印刷事業一般で必要とされている Glyph の粒度と行政現場で人名表記のために必要とされている Glyph の粒度が異なることは自明である。

下記の例では、龍の字について、Adobe Japan1 のコレクションでは、3 種類の Glyph が、Hanyo-Denshi のコレクションでは、5 種類の Glyph が登録されている^[5]。

<U+9F8D,U+E0106>に対応する字は、一般的には誤字とも考えられるが、姓名判断等により、敢えて一画欠損した字を用いる必要があった、とも判断しうる。



5. 異符号にすることの功罪

先に述べたように、符号化文字集合が異なれば、それぞれの包摂規準や統合規則も異なるものとなる。一般的には、使い分けが必要な文字には別符号を付与すればよい、と考えられがちであるが、必ずしもそうではない。

宮澤彰氏は、夙に

“比較的出しやすい結論は、細かく分ける方が情報量が多いのだから、分けておくべきであるという判断である。(中略)必ずしも情報量が多い方がよいというわけではない” “コード化にあたって、コード要素集合の要素を増やすことは、一般にコード化のコストを高くする。具体的には、コード表を人間が引くときの手間を増やし、校正の手間を増やす。たとえ自動的な手段を使っても、この点は変わらない。OCRでは認識できる文字数を増やせば、誤読率も高くなるのが普通である。” [6]

と述べているし、豊島正之氏も、

“文献は、解釈を経て釈文(解釈文字列)に翻刻され、その翻刻釈文が符号化されるのであって、当然、解釈者・翻訳の目的によって釈文も異なり、その結果は別の符号になる。” “実際の原文には、多くの字形衝突が生ずる(OCRはここで間違ふ)が、それらは釈文の解釈段階で殆ど解消され、一方、多くのバリエーションは統合・整理される。(衝突解消・整理できていないものは、そもそも「釈文」ではない)。”

“文字の符号化は、この様に一旦解釈された文字列が符号化されるのであって、原文が常に同一の符号列に変換される事は意味しない。” [7]と述べている。

用途によっては、印刷又は手書きの原文に異なる Glyph が現れていても、その解釈の方針によっては、同一の符号位置に統合して符号化する場合も、大いにあり得る。また、書籍などの編集方針によっては、“常用漢字に含まれる文字は、常用漢字の字体で、それ以外の文字は、表外漢字字体表の原則に基づいて”という場合も“常用漢字に含まれている文字も含め、すべての字体は、いわゆる康熙字典体とする”という場合もありうる。前者の原則に基づく符号化文字集合があったとして、それには常用漢字表に含まれる文字のいわゆる康熙別掲字形は不要だし、後者の場合にはそもそも常用漢字の字体は不要となる。

いささか 極端な例を挙げたが、符号化されるべき文字集合そのものが、言語文化的背景や使用目的によって変化する可能性を持つことは、お分かりいただけたと思う。

6. 用途による Glyph 粒度の相違

冒頭にも述べたように、ISO/IEC JTC1 が策定している規格群は、Character と Glyph をはっきり弁別する立場を取っている。その上で、単一もしくは複数の Glyph を情報交換用符号化文字集合の一つの符号位置に対応づけている。

この Glyph の側にも、複数の具体的な字形が同一の Glyph に属するか否かという判断が介在する。

ISO/IEC TR15285 の定義にも明らかなように、Glyph はいかなる specific design からでも独立した抽象的なものである。しかし、何を specific design と考えるかが、用途や立場によって異なる。

常用漢字表には、“(付)字体についての解説 第1 明朝体のデザインについて”として、同一の字体とみなすデザイン差についての記述がある。ここで述べられている《字体》と Glyph とが厳密に一対一に対応するかどうかについては議論の余地があるが、ここでは、《字体》と Glyph とを当面同一視して、改定常用漢字表に掲げられた例を用いて議論を進める。

下記の図は、「筆押さえ」等の有無に関する例である。この“明朝体のデザインについて”は、2000 年に答申された国語審議会答申「表外漢字字体表」にも踏襲されている。JIS X 0213 は 2004 年の改正の際、この趣旨に則り、不要な誤解と混乱を避けるために、表外漢字字体表に掲示されていた漢字については、規格票の例示字形を、“筆押さえ”があるデザインから“筆押さえ”のないデザインに変更した経緯がある [8][9][10]。

(4)「筆押さえ」等の有無に関する例

芝 芝 更 更 伎 伎
八 八 八 公 公 公 雲 雲

常用漢字表では、いわゆる“筆押さえ”を、デザイン差として、明確に同一の字体とみなしているが、一方、Adobe Japan 1 と呼ばれる Glyph Set では、“筆押さえ”のあるなしを別の Glyph として区別している。筆者が係わった最近の例でも、“汎用電子情報交換環境プログラム”では、“筆押さえ”のあるなしをデザイン差として同一視しており、“住基ネット統一文字”および“文字情報基盤整備事業”では、“筆押さえ”のあるなしを別の独立した Glyph とみなしている [11]。

66F4 更 更
E0100 E0101
Adobe-Japan1 Adobe-Japan1
CID=1995 CID=13441

また、中国の古典籍を読むための辞書として発展してきた従来の日本の漢和辞典とは一線を画する“新潮日本語漢字辞典”では、付録の“明朝体活字のデザインについて”という項目に、下記のような記述がある。

“また、デザインの違いであるとされている小異についても、小異のある字形を「別体」として積極的に収録した。これは実際の明朝活字で使用されている例が多く、実例を示すことが参考になると考えたからであり、これらの字形が常用漢字としてふさわしくないと考えたものではない。”

“なお、「明朝体活字のデザインについて」は従来拡大解釈されてきたきらいがある。たとえば、「更」は、当用漢字字体表以前の実際の明朝活字はすべて「筆押さえ」のついた「更」であったことは紛れもない事実である。しかし、これを常用漢字表の（付）でデザインの違いであるとされていることを理由に、「更」を旧字体の欄からはずしている例が多い。これは、過去の歴史から見て、あまり適切な取り扱いであるとはいえない。” [12]

この記述から透けて見える“新潮日本語漢字辞典”の編集者の考えは、常用漢字表（本辞典の発行時点では改定前のもの）の“明朝体のデザイン差について”の記述は尊重するものの、デザイン差はデザイン差として認識した上で、いわば“過去の歴史を踏まえたデザイン差は尊重すべきである”との立場を取っている。

このような立場を踏まえた上で、この“新潮日本語漢字辞典”を電子的に編纂する環境を想定すると、その環境の中では、“筆押さえのある”《更》と“筆押さえのない”《更》とを体系的に区別する必要があるが生じる。

7. VS の提案から IVD の実装へ

上述のように、情報交換用符号化文字集合では、単一もしくは複数の Glyph に対して固有な名前もしくはビット列を割り当てている。また、ある Glyph は抽象的な図形概念であり、デザイン差を含むある粒度での具体的視覚表現を包含している。

しかし、ある Glyph が包含する具体的視覚表現の範囲（粒度）は、言語文化によっても、使用目的によっても、変化する。

筆者らが、1995 年に提案した Variation Selector（以下、VS と略記する）というメカニズムは、ある符号位置に統合される Glyph の相違を、規格適合性を満たした上で再現する方法としては、非常に有効であったと考えられる。この VS のメカニズムは、ISO/IEC 10646 でも数学記号や古モンゴル文字などに適用されている [13]。

一方、筆者らが当初想定していた VS の用い方は、統合漢字の CTJK 各欄に掲示されている各国の原規格の例示字形を区別して再現するためのものであったが、6. で述べたように、Glyph の粒度に対する要求そのものが、言語文化や用途によって異なるために、国際標準として一律の粒度で Glyph の異同を判断することには、さまざまな軋轢が予想されたため、長く、CJK 統合漢字への VS の適用は留保されてきた。

Unicode Technical Standard #37 “Ideographic Variation Database” [14]（以下、IVD と略記する）は、この Glyph 粒度の言語文化や使用用途による差異を認識した上で、VS メカニズムを適用するために、使用者コミュニティごとに漢字の符号位置と VS との組（Ideographic Variation Sequence, 以下 IVS と略記）をコレクションとして登録し、複数コミュニティ間での IVS の衝突を避ける運用規定である。

IVD は、2005 年に Unicode Consortium の Technical Standard として制定され、ほどなくこの TS を規格内で引用することにより、ISO/IEC 10646 においても正規の規格となった。

IVD へは、2007 年に Adobe Japan1 [15]、2010 年に Hanyo-Denshi [16] という二つのコレクションが登録されている。

また、この IVD の利用環境については、Adobe 社の Acrobat 9 を嚆矢に、Microsoft 社の Windows 7、Apple 社の OS X など、順次実装が進んでいる。

8. 何を区別し何を区別しないか

符号化文字集合における、Character と Glyph の二つの層を軸に、Character もしくは Glyph の異同判断が、一般的には、言語文化の相違や使用目的の相違の影響を受ける、いわば、状況依存的なものであることを述べてきた。

本項では、上述の議論を踏まえた上で、現行の JIS 漢字コード 及び UCS で、技術的にどのような区別が可能かを、具体的な例と共に示す。

a) JISX 0208 において、別符号が付与されている例

下記の例は、JIS X 0208 において、それぞれ独立した符号位置が付与されているものである。これらは、単に Glyph が相違するというだけではなく、情報交換用符号化文字としても、独立している。

檜山 ⇄ 桧山

鰺沢 ⇄ 鰺沢

富田 ⇄ 冨田

b) JIS X 0208 と JIS X 0213 で包摂の範囲が異なる例

下記の例は、あまりに有名であるが、JIS X 0208 の 1978 年版と 1983 年版とで、同一符号位置（10 区 18 点）に例示されている字形が変更された。1997 年版で、包摂規準が導入された際、同一符号位置に包摂するものと明示された。その後、JIS X 0212（補助漢字集合）でも、JIS X 0213（拡張漢字集合）でも、別符号位置が付与され、情報交換用符号化文字としても、独立するものとなった。

鷗外 ⇄ 鷗外

c) JIS X 0213 の 2000 年版と 2004 年版で字形が異なる例

下記の例は、表外漢字字体表との整合性を取るため、JIS X 0213 が 2004 年に改正された際、表外漢字字体表に掲示されている字形に合わせて JIS X 0213 の例示字形が変更されたものである。これらについては、JIS X 0208 においても、JIS X 0213 においても、包摂規準によって明確に包摂されているが、一般社会慣習上、地名や人名としては、区別に拘る場合が多くあり、喧しい議論的となった。

Microsoft 社の Windows が VISTA 版で、全面的に JIS X 0213:2004 に対応した際、ある種の社会現象を引き起こしたことも記憶に新しい。

葛城市 ⇄ 葛飾区

辻 ⇄ 辻

d) JIS X 0213 の包摂規準と UCS の統合規則が異なる例

下記は、JIS X 0208、JIS X 0213 の包摂規準では包摂され、ISO/IEC 10646 では統合の対象外として、独立した符号位置を与えられているものである。それぞれ事情が異なるので、個別に説明する。

《高》《高》の例：《高》には、当初の UCS 開発時の“原規格分離原則”により、《高》の字とは別の独立した符号が与えられている。

高田 ⇄ 高田

《吉》《吉》の例：《吉》は、CJK 統合漢字拡張 B に採録されている。筆者の私見であるが、この二つは、CJK 統合漢字の統合規則（Annex S）に照らしても、明らかに統合されるべきものであるが、拡張 B 開発時の作業上の瑕疵によって別符号化されたものであると考えられる。

吉田 ⇄ 吉田

《靖》と《靖》の例：《靖》は、CJK 互換漢字として独立した符号位置を与えられている。CJK 互換漢字は、本来統合の対象となる漢字で何らかの理由により、過去の規格や実装との互換性を保つために符号化されたもので、UCS でも Unicode でも、その利用に当たっては、慎重を期することを求めている。さらに、Unicode では、Normalization forms[17]と呼ばれる Standard Annex で、対応する統合漢字への変換を規定している。そのため、通信経路で予期せぬ統合漢字への正規化が行われる危険があり、この点も含めて、使用に当たっては慎重を期する必要がある。

また、現在、Unicode Consortium が積極的に IVD の利用を推奨している論拠の一つとして、この Normalization プロセスによる互換漢字の統合漢字への正規化のリスクが挙げられている。

靖 ⇄ 靖

e) IVD で Glyph 区別が出来る例

下記の例は、JIS X 0213:2004 において字形変更されたものであるが、JIS の符号化文字集合としては、包摂規準により、区別して扱うことは出来ない。しかし、IVD 登録することにより、IVS のメカニズムで Glyph としての区別が可能となる。その際、Glyph としての区別の粒度が、IVD に登録された Glyph コレクションによって異なることに留意されたい。

葛城市 ⇄ 葛飾区

辻 ⇄ 辻

丈 ⇄ 丈 (Hanyo-Denshiでは区別しない)

9. 区別することの必要性を明確化することの必要性ということについて

8. で示したように、IVD の導入により、同一符号位置に統合される漢字でも、IVD に登録された IVS を駆使することで、Glyph の相違を表現することが可能となっている。さらに、Glyph を区別する粒度をどのように設定し、Glyph の異同をどのように線引きするかは、IVD の登録主体となるユーザーのコミュニティに委ねられている。

このような技術的環境が国際標準として実現した現在、情報交換用符号化文字集合標準の利

用者に求められるのは、利用者自身が、どのような目的のために、どのような情報の保存や交換を、どのような粒度で行いたいかを、明確に意識化し、要求仕様ないしは凡例として文書化することである。

10. あとがき

筆者は、1998年に発行された『電腦文化と漢字の行方』に寄せた一文を、下記のような言葉で結んでいる。

“規格が社会的な機能を全うするためには、変化を変化として認めたい。規格に対する要求を明確な社会的合意として形成していくための、積極的で建設的な議論が、なによりも大切なことなのだ。” [18]

この考えは、10余年たった現在に至るまで、全く変化していない。そして、筆者を含め、要求される側の符号化文字集合を開発する立場にある技術者集団は、それぞれに、要求する側の声に真摯に耳を傾ける努力をしつつ、少しずつそれに応えるための標準開発を行ってきた。

IVDは、標準開発者側からの、一つのささやかな回答である。

その回答では、はからずも、技術側で何らかの判断を行うことなく、利用者側に判断を委ねることとなった。技術側は、文化的多様性をそのまま受け入れたうえで、多様な文化が共存できる技術的機能を提供する道を選んだのだ。

利用者集団が、それぞれの文化的背景と目的に応じて、この機能を十全に活用して下さることを祈念している。

VSの標準化からIVDの実運用に至る道筋では、多くの方々が主体的に関わり貢献されてきた。この方々のお名前を差し置いて、2010年に急逝した畏友故樋浦秀樹氏への感謝の気持ちを表明することをお許しいただきたい。

11. 参考文献

[1] ISO/IEC 10646:2003 Universal Coded Character Set

[2] ISO/IEC 10646:2003 UCS Annex S (Informative) “Procedure for the unification and arrangement of CJK Ideographs”

[3] ISO/IEC 10646:2003 UCS CJK Unified Ideographs Block

[4] JIS X 0213:2004 7ビット及び8ビットの2バイト情報交換用符号化拡張漢字集合

[5] Ideographic Variation Database
“Combined registration of the Hanyo-Denshi collection and of sequences in that collection”

<http://www.unicode.org/ivd/data/2010-11-14/>

[6] 宮澤彰 「書誌情報データベースから見た文字コード」 『インターネット時代の文字コード bit 別冊』 所収、共立出版、2001 年刊

[7] 豊島正之 「JIS 批判の基礎知識」 日本語の文字と組版を考える会 1998 年 3 月 8 日 予稿
www.joao-roiz.jp/mtoyo/on-JCS/mt-gene.pdf

[8] 常用漢字表 昭和 56 年 10 月 1 日 内閣告示 第 1 号

[9] 表外漢字字体表 平成 12 年 12 月 8 日 国語審議会答申

[10] JIS X 0213:2004 前掲

[11] Ideographic Variation Database
Combined registration of the Adobe-Japan1 collection and of sequences in that collection
<http://www.unicode.org/ivd/data/2007-12-14/>

[12] 新潮日本語漢字辞典、新潮社、2007 年刊

[13] ISO/IEC 10646:2003 Universal Coded Character Set 20.4 Variation selectors

[14] Unicode Technical Standard #37
UNICODE IDEOGRAPHIC VARIATION DATABASE
<http://unicode.org/reports/tr37/>

[15] <http://www.unicode.org/ivd/data/2007-12-14/>

[16] <http://www.unicode.org/ivd/data/2010-11-14/>

[17] Unicode Standard Annex #15
UNICODE NORMALIZATION FORMS
<http://unicode.org/reports/tr15/>

[18] 「要求する側の責任ということについて」 『電腦文化と漢字の行方』 所収（平凡社、1998 年刊）