文字集合の独立性

今回は、ちょっと厄介で理屈っぽい話になる。じっくりお付き合いいただければ幸い。

結論を先に述べておくと。

二つの文字集合があったとき、それぞれの文字集合は、互いに共約/通約不可能である。

ここで、共約/通約不可能と書いたのは、英語のincommensurableの訳語のつもり。

この言葉自体、このブログをお読みの方々には、耳新しいかもしれない。

元来は数学の世界で、互いに通分できな分数の関係を表す言葉だったらしい。

科学史家のトマス・クーンが、「体系、概念、方法論などに違いを持つ異なる体系（パラダイム）同士の間で、概念間の対応付けがうまく出来ない状態のことを指す」（ウィキペディアから引用）のに使ったのが始まり。

例えば、古典的なニュートン力学と、相対性理論の一部として取り込まれたニュートン力学とでは、互いに通約不可能であるといったふうに用いる。

ぼくが、この問題を考えるようになったのは、文字情報基盤のレパートリーをIVDに登録する作業を行っていたころのことだ。

文字情報基盤整備事業は、汎用電子情報交換環境整備プログラムの後継事業としての性格があり、基本的には、汎用電子情報交換環境整備プログラムの成果を継承している。しかし、汎用電子情報交換環境整備プログラムがどちらかというとアカデミズムよりの調査研究プロジェクトとしての色合いが強かったのに対して、文字情報基盤整備事業の方は、地方自治体などでの利用を前提とした現実に即した対応をするという側面が強い。そのため、細部で、厳密には（学問的には）汎用電子情報交換環境整備プログラムの通りなのだけれど、実際にはそうも言っていられないのだよね、といった変更・修正が行われた。それが、いざ、文字情報基盤のコレクションをIVDに登録する際に、問題となった。

ぼくの名前、龍生の《龍》の字を例に説明しよう。

《龍》の字は、《邊》《邉》に次いで、異体字が多いので有名だ。

汎用電子コレクションには、11種類が、文字情報基盤コレクションには6種類が登録されている。このうち、IVSを共有するものが5種類、汎用電子のみに登録されているものが6種類、文字情報基盤のみに登録されているものが1種類ある。





この二つの図を見比べるだけで、汎用電子情報交換環境整備プログラムと文字情報基盤整備事業の目的の違いなどが透け見えて興味は尽きないのだが、ここでは深追いしない。

文字情報基盤コレクションをIVDに登録する際に問題となったのは、字形が非常に似ている汎用電子のVSを文字情報基盤のVSとして流用するかどうか、という点だった。UCSの符号位置もやはり限りある公共資源なので、出来れば無駄遣いはしたくない。

ぼくは、この議論があったJSC2の会議の場を今でもよく覚えている。ぼくの主張は、字形が類似しているからと言って、汎用電子における字形と文字情報基盤における字形が、同じ字体を指し示している保証はない、というものだった。ぼくの主張は、ある例示字形が指し示している字体の範囲は、字体相互の関係の上で相対的に定まるものなので、字体集合が異なれば、たとえ同じ字形でも、その字形で示される字体の範囲は異なる、というものだった。

前回用いた思考実験で考えてみよう。

まず、百均ショップで買ってきた塩ビ製の透明な小袋を用意する。

この小袋に、汎用電子コレクションの場合は、E1013(JA4622)、E1014(JTC0A9)、E0105(IB3312)、E1016(JTC0A8)、E1017(FT2861)、E0108(TK01101760)、E0109(TK01101790)、E010A(TK01101810)、E010B(TK01101820)、E010C(TK01101830)、E010D(TK01101840)という符合と、対応する字形（IVDコレクションのPDFから切り取ってきた例示字形）を貼り付ける。

その上で、これら11個の小袋に、そこらへんに散らばっている雑誌や新聞、名簿や古文書から、龍の字を切り取ってきて、ラベルの字形が一番似ていると思われる小袋に入れていく。例えば。



右下の《テ》の部分に注目しても、E0104、E0105、E0109、E010A、E010B、E101C、E010Dなどが候補として挙げられよう。

他方、文字情報基盤コレクションの場合は、どうだろう。

候補は、ぐっと減って、E0104とE1015に絞れそうだ。

例に挙げた手書きの墨文字が、それぞれどの小袋に入るかは、最終的には、当事者（文字を書いた人か、この字形を使って名刺を作った本人とかに選んでもらうしかないのだが（固有名詞と字体の関係は、次回のアーティクルで議論する）、今考えている問題は、文字情報基盤のE0104(MJ030125)とE0105(MJ030127)が指し示す字形の集合としての字体の範囲と、汎用情報コレクションのE1014(JTC0A9)、E0105(IB3312)が指し示す字形の集合としての字体の範囲が、同じかどうか、ということ。

ぼくの立場は、明確に、異なる。だって、汎用電子コレクションには他に、E0109(TK01101790)、E010A(TK01101810)、E010C(TK01101830)、E010D(TK01101840)といった似たようなラベルを貼った小袋があるではないか。

今の試行実験で、問題圏が見えてきただろうか。Unicode ConsortiumのIVDについての説明文書には、glyphic subsetという言葉が出てくる。なかなかうまい言い回しなので、流用させてもらって説明する。

そもそも、IVSのアイディアは、通常は一つの符号位置（＝字体）に統合されてしまう複数の字形を区別するために提唱された。当時は、例えば有名な《骨》など、日本、中国、韓国、台湾で、符号位置としては統合されているが、それぞれの国や地域で伝統的に用いられている字体に明らかな差異が認められる符合位置に適用することが念頭に置かれていた。

しかし、統合漢字へのIVSアーキテクチャの適応は、Unicode ConsortiumがIVDというIVSコレクションの登録制度を提唱するまで、実現しなかった。この辺りの委細は、「ユニコード戦記」にすでに書いた。

いずれにしても、異なるIVSに対応付けられる複数の字形があるとき、base characterが同じなら、これらの字形は、UCSとしては、同じ符合位置≒字体に統合されている。

UCSとしては、同一字体なのだが、IVD的には区別されているわけだ。IVDドキュメントを書いたKen Lundeのglyphic subsetという概念は、UCS的には一つのglyphに対応付けられる字形の集合のうち、あるIVSに対応付けられる字形の集合を指しているわけだ。



たとえば、上記左のような筆文字の《龍》の字形があって、これを、例の字体毎の小袋に入れることを考える。

MJコレクションの場合は、それほど迷うことはないだろう。多分、E0104の袋。

では、汎用コレクションの場合は、どうだろう。$0104でもいいが、E0109やE010A、はたまたE010Cの袋に入れてもいいような気もする。

問題は、一旦汎用コレクションのE0109やE010A、E010Cなどの袋に入れた、この字形を、MJコレクションの袋に入れ替える時に起こる。E0104の袋に入っていれば、そのままMJコレクションの中のE0104の袋に入れ替えればいいのだが、E0109や、E010A、E010Cの袋に入っていたら、行き先がなくなってしまう。

逆の場合もある。MJコレクションでE0104の袋に入っているいくつもの字形を汎用コレクションの袋に移すことを考えると、汎用コレクションのE0104の袋に移すことは出来ない。E0105やE010A、E010Cのどれかどころか、他のE0103からE010Dまでの11個すべての中から、改めて一番近そうな袋に入れ直さなければならない。

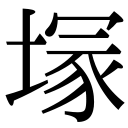
このことこそが、Kenの謂を借用すると、glyphic subset間のincommensurabilityということになる。すなわち、MJコレクションにおけるU+9F8D\_E0104と汎用コレクションにおけるU+9F8D\_E0103は、incommensurableなのだ。

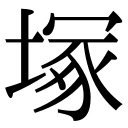
先を急ごう。このincommensurability問題は、IVDにおけるglyphic subsetだけではなく、異なる字体集合間すべてに存在する。

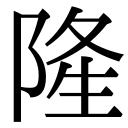
例えば、JIS X 0208とJIS X 0213。ざっくり言ってしまうと、単独のJIS X 0208とJIS X 0213の一部としてのJIX 0208とは、incommensurableなのだ。そう、かつて恩師村上陽一郎先生が例として引かれた、古典力学としてのNewton力学と、相対性理論を前提としたNewton力学との関係と全く同型の関係が、ここには存在する。

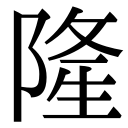
本協議会の副会長でもあるアドビの山本太郎さんやJISの組版規則やJLreqの執筆者小林敏さんを初めとする多くの人たちが、JIS X 0213策定のころ懸念していた包摂分離の問題とは、つきつめると、この異なる文字集合間のincommensurabilityの問題だったのだ。

具体例で示そう。

1-36-45、U+585A

 1-15-55、U+585A

 1-46-20、U+9686

 1-93-61、U+9686

じつは、《つか》のケースと《たかし》のケースで、Unicodeが絡んでくると、若干事情は異なるのだが、ここでは、JIS X 0208とJIS X 0213との関係だけに注目しよう。

まずは、《つか》。

手元に、文化庁文化部国語課が、当時の国語審議会における審議に資するために復刻したJIS C 6226:1978がある。これを見ると、36区45点の字形は、《塚》の形に近い。

一方、同じ資料に含まれているJIS X 0208:1983では、36区45点の字形は、《塚》の形に近い。そして、JIS X 0298：1997。この版の、36区45点の記述が面白い。過去の規格票の字形（参考）の欄に、78として《塚》の字形、78/4Xとして《塚》の字形が掲げられている。そして、JIS X 0208:1990から加えられた包摂規準では、129)として、いわゆる豕の点があるものとないものを包摂している。（ぼくの手元には、JIS X 0213:1990がないので、JIS X 0213:1997を参照している。）

そして、JIS X 0213:2000。

ちょっとやっかいな話だが、JIS X 0213には、JIS X 0208のすべての面区点位置が含まれている。表層的には、JIS X 0213は、JIS X 0208の区点位置に、新たにいくつかの面区点位置を追加した形となっている。

しかし、JIS X 0213の規格票には、面区点位置の一覧表部分にJIS X 0208相当の例示字形が掲載されているが、附属書6（規定）漢字の分類及び排列には、追加部分だけが記載されていて、JIS X 0208相当部分への詳説はない。

《隆》と《隆》についても見ておこう。

こちらの方は、JIS X 0208では、一貫して《隆》の方の字形が掲載されている。

その上で、包摂規準のd)連番132で、包摂されることが明記されている。

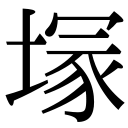
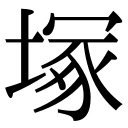
そして、JIS X 0213。《隆》に新しい面区点位置を与えた上で、包摂規準の詳説については、JIS X 0208とほぼ同等のものを掲げ、連番133のところに、「（この規格では適用されない）」という記述がある。いまさら、このような細部をほじくり返しても、なんだかなあ、という気がするので、これ以上深追いしないが、何だか釈然としない。

JIS X 0208で包摂されていながら、JIS X 0213では、別の面区点位置を与えていた例は、他にもいくつもある。

いよいよ、ぼくの主張の本題に入ることにしよう。

根幹となる主張をもう一度確認しておこう。

文字集合が異なれば、ある同一の字体名に含まれる字形の集合は互いに共約不可能である。

すなわち、JIS X 0208を単独の文字集合として見たときの区点位置36-45と、JIS X 0213全体の文字集合の一部として見たときの面区点位置1-36-45では、そこに掲示してある例示字形が同一だとしても、その区点位置または面区点位置に帰属する字形の集合は別のものなのだ。に類似した字形（よりも一画多い）は、JIS X 0213では、別の面区点位置1-36-45に帰属する。

JIS X 0213とUnicodeの間、Unicodeで統合分離が行われる前と後との間でも、事情は全く変わらない。

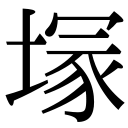
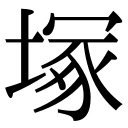
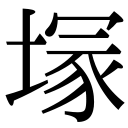
以上述べてきたことは、次のように言い換えてもいいだろう。

ある字形が対応付けられる字体は、その字体を要素として含む字体集合が異なれば、たとえ同じ文字名を持った字体であっても、そこに帰属する字形が同じであることは保証されない。

もう一つ、別の言い方。

符号化された文字列（字体名の列）が送信者側と受信者側で同一であることを保証するためには、送信者側と受信者側が同一の文字集合を用いなければならない。

当たり前と言えば当たり前の話だが、一見同じ字体集合に見えても、改正等で包摂分離や統合分離が行われると、それらの字体集合は互いに共約不可能となり、通信における同一性は保証されなくなる。

今回は、かなり煩瑣な議論になってしまった。しかし、実際の社会生活においては、ここまで厳密な運用は事実上不可能だろうし、大きな不都合が生じるとも考えにくい。ぼく自身、以前非常勤で出講していた大学で、大学当局から提供された受講生名簿と学生本人から提出されたレポートの氏名がとで異なっており（学生の方が）成績処理の際、いささか混乱した程度の経験しか無い。とはいえ、このような稀な不都合に遭遇した際には、背後に上記のような根源的な問題があることを思い出していただくことで、対処の一助になれば幸い。