起こるべくして起こった？　マイナンバー紐付け過誤問題

神奈川県保険医協会　医療情報部　部長　藤田倫成

はじめに

多くの読者諸氏は、御自身のデジタル機器がＷｉｎｄｏｗｓ・Ｍａｃ／ｉＰｈｏｎｅ・Ｕｎｉｘ／Ｌｉｎｕｘ、Ａｎｄｒｏｉｄといった違いは有れど、画面で「漢字」が表示されるのは「当たり前」と思われて居られる事だろう。だが、それはユーザーの目に触れないダケで、裏に涙ぐましい歴史と改良の変遷が有った事は、一部の趣味人等を除き、あまり知られていない。しかし、そこにこそ今問題となっている保険証のマイナンバーによる資格確認の際に発覚する紐付けに失敗する大きな一因が内在している。

文字コード体系の変遷

※バイトコード

古くは電話線を介して電動タイプライターの遠隔接続をしたテレタイプが起源と考えられ、それを初期のデジタルコンピューターの入出力に流用したという歴史が有り、この時には英文アルファベット大文字と数字と若干の記号と改行とかの制御記号のみであった為 32 文字(5bit)有れば事足りた。その後アルファベット小文字や記号が追加されて 64 文字(6bit)、記号部分の拡張と通信時の間違いを検出する為のパリティ 1 ビットを加えた 128 文字(7bit&8bit)が生まれ、パリティの代わりに図形等を割り当てた 256 文字(8bit=1byte)のコードが派生し、それぞれが工業規格として IEEE を経て ANSI（後の ISO）に纏められた。この時の ANSI 規格が日本の JIS 規格として準じて制定され、日本でも使われるようになった。しかし、バイト(8bit)コード後半の図形区にカタカナ（半角）を割り当てた 8bitJIS は日本独自規格であり、世界的にはローカルなものに過ぎず、勿論ちゃんとした漢字に対する割り当てなぞ未だ存在しなかった。

※ワード（２バイト）コード（JIS 第１水準＋第２水準＝区点 9999 文字への拡張）

そして 1978 年に旧 JIS C 6226-1978（現 JIS X 0208 相当）の第１水準、第２水準が規定された。第１水準には漢字 2965 字（常用漢字 1945 字とその他の人名用漢字）、第２水準には第１水準よりも使用頻度の低い地名や人名・一部旧字体などを含む 3384 字が含まれ、この第１水準＋第２水準の 6349 字にかな、英数、特殊記号など非漢字 453 字を加えた合計で 6802 字となる。この後、1983 年に

第１水準漢字 2965 字は据え置きで第２水準が 3388 字と４文字増え、非漢字 524 字と合わせて 6877 字となった。1990 年に第１水準漢字 2965 字は据え置きで第２水準が 3390 字へと２文字増え、新たな補助漢字を含む非漢字 524 字と合わせて 6879 字となった。1997 年改定では一部の入れ替えを除き文字数は変わらず、第１水準 2965 字＋第２水準 3390 字＋非漢字 524 字で 6879 字のまま、そしてこれ以降は第１＋第２水準の 6355 字は変わらない。

※多バイトコードへ

2000 年、旧来の第１水準＋第２水準＝6355 字の漢字に第３水準 1249 字、第４水準 2436 字が新たに追加され、そして非漢字も 1183 字に増えて 11223 字となった。2004 年改定では第３水準が 1259 字に増え（他は据え置き）全部で 11233 文字となり現在に至る。

しかしながら当時のＰＣ等で普及する一般的な OS ではワードに収まる第２水準までの実装に留まり、第３水準以降については μTRON 等の特殊な OS 環境を除き、標準実装されているものは寡聞にして知らない。この結果、使う端末毎、若しくは事業所単位とかの小グループでのみ共通化された、全国的には互換性の無い（第３及び第４水準に存在する漢字も含む）ユーザー定義文字（所謂「外字」）セットが多数生まれる事となった。

※Unicode へ

これらは単一バイト(8bit)が表せる文字セットがシフトイン等のコードページ操作をしない場合、最大 256 文字しか扱えず、欧米の幾つかの国で必要となったウムラウト等の限られた文字での英語圏アルファベット 26 文字を変化させたローカル文字の導入とは異なり、漢字等の字体の多い文字が 1byte=8bit=256 字に収まらない事から、多バイト文字と呼ばれる多数の文字セットを持つ複数の言語の表現に互換性が期待できない事になる。そこで 1980 年代にゼロックスが提唱し、多くのベンダーが参加するユニコードコンソーシアムにより策定されたのが Unicode であり、国際規格の ISO/IEC 10646 と Unicode 規格は同じ文字コード表になるよう協調して策定されるようになった。

残念ながら 1980 年代当初の構想の Unicode は 16 ビット固定長で、1Word = 2byte = 16bit = 2 の 16 乗 = 6 万 5,536 個の符号位置に必要な全ての文字を収録しようとする想定は Unicode 1.0 公表後、拡張可能な空き領域２万字分を巡り、中国、日本、台湾、ベトナム、シンガポールの追加漢字約１万５千字、古ハングル約５千字、未登録言語の文字など想定以上の文字追加要求が起こる事となった。こうして Unicode の、16 ビットの枠内に全世界の文字を収録するという計画は早々に破綻し、1996 年の Unicode 2.0 の時点で既に、文字集合の空間を 16 ビットから広げることが決まった。この時、それまでの 16 ビットを前提としてすでに設計されていたシステムをなるべくそのままにしたまま、広げら

れた空間にある符号位置を表現する方法として、特殊な方法（サロゲートペア、代用対）が定義されるなどした。

日本では2000年に第３水準＋第４水準(JIS X 0213)が制定された（前出）が、この際、新たに採用された文字でUnicodeに無かったものの一部は、第0面のBMP（Basic Multilingual Plane、基本多言語面）に収録できず、第2面への収録となった（Unicodeが最終的にJIS X 0213への対応を完了したのは2002年）。このため、JIS X 0213収録文字をUnicodeで完全にサポートするには、追加漢字面をサポートしたOS、フォント、アプリケーションが必要となり、第２水準までしか扱えないShift_JISなど、Unicodeで規定される以外のエンコーディングを利用する場合であっても、JIS X 0213に対応するフォントやアプリケーションが必要となり、これが汎用端末でのJIS第３水準、第４水準普及の足枷、そして汎用性の無い外字セットの乱立の元となったと考えられる。

Unicodeは1991年10月にUnicode 1.0.0、収録文字数7,161（JIS X0201）を制定して以来、何年かに１度、年によっては年内に複数回の改定を経て、最新は2022年のUnicode 15.0.0、収録文字数149,186（ISO/IEC 10646:2022）に肥大化している。この中では1996年のUnicode 2.0.0、収録文字数38,950に増えた際のハングルの大移動等の大幅改変でUnicode1.xとの互換性を失うなど、「Unicodeの規格全体に付けられたバージョン」の他に「Unicodeを構成する個々の要素（BMP,SMP,SIP,TIP,SSP,NOR）の規格に付けられたバージョン」も存在し、その違いによって一部の互換性が「保証されない」仕様となっている。これらは、マイナ保険証の確認画面等で未定義文字（ちゃんとした漢字が表示されず、代わりに俗に言う「下駄文字」と呼ばれる太い横２本線や中点や斜線などが表示される）の発生する一因ともなっている。

※日本独自の各行政内漢字統一コードの乱立と縮退の方向性

先述の様に、汎用端末でほぼ互換性が保たれるのは、JIS第１水準＋第２水準の漢字及びかな、英数、（初期バージョンでの限られた）特殊記号などの非漢字の一部に限られ、他の文字を扱おうとすれば、互換性の無い（第３及び第４水準に存在する漢字も含む）ユーザー定義文字（所謂「外字」）を使わない限り、日常の業務が遂行できない事態となった。この為、使う端末毎、若しくは事業所単位とかの小グループでのみ共通化された、外字セットが乱立する結果となった。

国もその事態は認識しており、実際に経産省の下の独立行政法人情報処理推進機構（IPA）が文字情報基盤整備事業として文字情報基盤の文字（Moji Joho Kiban Ideographs、文字情報基盤の文字、MJ文字）として戸籍統一文字、住基統一文字の58,712文字を整理して2017年にISO/IEC 10646第５版として規格化が完了した（2020年にIPAから一般社団法人 文字情報技術促進協議会へと民間移行された）。この文字情報基盤整備事業では、MJ文字情報一覧表に示す約６万字の文字情報基盤の文字（MJ文字集合）と、約１万文字のJIS X 0213（JIS第１水準～第４水準）の文字との対応関係を整備

しているが、縮退先を「複数」示したり、縮退先を「示さない」漢字が存在し、結果として縮退先が無いものについては読み仮名に置き換えるか、複数の文字からなる熟語等に置き換えるといった対応が必要となり、更なる混乱の一因ともなっている。

また、UCS(Unicode)と住基文字の一部では、同じプログラム（計算式）を使っても、文字コード間の互換性（変換）が一対一に保てない領域が幾つか有り、そういった文字が使われていた場合、端末に正しい文字が表示されるという保証に期待はできない。

成果として公表されている MJ 文字情報一覧には、MJ 文字コードのほか、Unicode（UCS）、JIS X0212、X0213、登記統一文字番号、戸籍統一文字番号、住基ネット統一文字コード、入管正字コード、入管外字コード、そして紙媒体の漢和辞典等（大漢和、日本語漢字辞典、新大字典、大字源、大漢語林）の参照する文字番号が記載されている。

少し例を挙げてみよう。MJ 文字コード「MJ000001」で表される文字は「ＸＸ寧々」さんとかで用いる「々」の文字である。この文字の UCS(Unicode)は「U+3005」、住基ネット統一文字コード「J+AD1D」、入管正字・外字コード無し、X1213「1-01-25」、登記統一文字番号は無し、参考としての読み「おなじ・くりかえし・のま」、大漢和「97」、日本語漢字辞典は無し、新大字典「110」、大字源「43」、大漢語林は無しである。我々に馴染みの深い「医師」の「医」で見てみると同じ様に見える２文字が在り、MJ 文字コード「MJ007820」、UCS(Unicode)「U+533B」、戸籍統一文字番号「030420」、住基ネット統一文字コード「J+533B」、入管正字・外字コード「0x533B」、X0213「1-16-69（常用漢字＝第１水準）」、登記統一文字番号「00030420」、参考としての読み「イ・エイ・いやす・くすし」、大漢和は無し、日本語漢字辞典「1097」、新大字典「1435」、大字源「10449」、大漢語林「1074」の文字と、MJ 文字コード「MJ007821」、UCS(Unicode)「U+533B」で同じ、戸籍統一文字番号「030470」、住基ネット統一文字コードは無し、入管正字・外字コードは無し、X0213「1-16-69」で同じ、登記統一文字番号「00030470」、参考としての読み「イ・エイ・アイ・うつぼ」、大漢和「2680」、日本語漢字辞典は無し、新大字典「1466」、大字源無し、大漢語林無しが存在する。ちなみに旧字体の「醫」にも「MJ026547」「MJ026548」の２文字が存在し、UCS(Unicode)「U+91AB」で同じ、戸籍統一番号「454010」「454140」、住基ネット統一文字コード「J+91AB」「無し」、入管正字コード「0x91AB」「無し」、X0213「1-78-48（第２水準）」で同じ、登記統一文字番号「00454010」「00454140」、読み「イ・エイ・いやす」で同じ、大漢和「40006」「無し」、日本語漢字辞典「1098」「無し」、新大字典「17608」「無し」、大字源「無し」「10448」、大漢語林「1075」「無し」である。尚、異体字の「毉」は「MJ014933」、UCS(Unicode)「U+6BC9」、戸籍統一文字番号「189340」、住基ネット統一文字コード「J+6BC9」、入管正字・外字コード「0x6BC9」、X0212「38-22」、X0213「2-78-08（第４水準）」（以下略）の１文字だけである。

戸籍→住基→個人番号（マイナンバー）への乗り換えと互換性

※住民基本台帳のネットワーク化

住基ネットとは、平成 11 年住民基本台帳法（住基法）の改正により、地方公共団体共同のシステムとして各市町村の住民基本台帳のネットワーク化が図られたものである。

セキュリティ対策としてなのか、収載情報の収納されるフィールドやレコード構成、文字コード等の技術資料が公式には一般に公開されていないので、分かる範囲での推測とならざるを得ないが、住基ネットが管理する情報は、平成 20 年 3 月 6 日の最高裁判所判決（合憲判決）の「○ 住基ネットによって管理、利用等される本人確認情報のうち、４情報（氏名、生年月日、性別及び住所）は、人が社会生活を営む上で一定の範囲の他者には当然に開示されることが予定されている個人識別情報であり、個人の内面に関わるような秘匿性の高い情報とはいえない。（法令に基づき必要に応じて他の行政機関等に提供され、事務処理に利用されてきたもの）○ 住民票コードは、住基ネットによる本人確認情報の管理、利用等を目的として、都道府県知事が無作為に指定した数列の中から市町村長が一を選んで各人に割り当てたものであるから、上記目的に利用される限りにおいては、その秘匿性の程度は本人確認情報と異なるものではない。」とされている事から、基本４情報（氏名・住所・生年月日・性別）と住民票コード及びこれらの変更情報に限定されているであろう事が読み取れる。

※住基ネットにおける揺らぎ

そして、性同一性障害者の性別の取扱いの特例に関する法律（通称「GID 特例法」、平成１５年７月１６日法律第１１１号）による戸籍の性別を変更する取扱いが定められた時よりも施行されたのが昔である事から、戸籍上の後日改変された性別や、外国人も含む通称、そして旧姓といった情報について考慮されているとは考え難く、住基ネットは旧時代のシステムと言わざるを得ないだろう。

そして、もう一つの問題が氏名や住所の「読み」等である。これは現在の戸籍法には記載されておらず、2023 年 2 月 2 日の法制審議会の部会で氏名の読み仮名を新たに戸籍に加わえる審議がされた事からも明らかである。2024 年に海外で利用が始まるマイナンバーカードへのローマ字表記や、行政手続きのデジタル化促進のため、読み仮名もカタカナで戸籍に追加するとの説明から、逆に旧来の戸籍を基にした住基ネット、そしてそれを利用した現行のマイナンバー制度には「読み」や外国人の氏名表記における、アルファベットなのか、カタカナによる音表記なのか、そして通称についての扱いがどうなっているのかといった点についても統一されていない事が類推されよう。現に私の診療所を受診されている患者さんで、今の保険証はアルファベット大文字のみで、スペースで区切られた４パートの

氏名表記がされているにもかかわらず、横浜市の乳幼児医療証は中点「・」で区切られたカタカナ表記で、更に終わりまで全てが記載されていない方も居られる。（医療証については記載スペースが無くなったダケで、元データとしては最後まで存在しているのかもしれないが）

海外の方のカタカナによる音表記についても、例えば「金」を「キム」「キン」と読むとか、「ホワン」「ホン」「ホワング」の様にカタカナ化する際に揺らぎを生じたり、区切りに全角及び半角スペース「 」、全角及び半角中点「・」、全角及び半角ハイフン「ー」、全角及び半角マイナス記号「－」、全角及び半角等号「＝」を用いる等、統一されていない様に見える。そしてハングル等の外国の文字については、先出の Unicode のバージョン問題もあり、使用するシステムによっては今の日本の漢字区に重複して存在した時期もあるため、表示されるべき正しいコードが有っても正しく表示されないか、漢字が別の文字に表示されるといった問題も当然予見されなければならない。

※職権によるマイナンバー参照

今日のデジタル庁ＨＰでも、マイナンバーカードの取得は任意であると記載されている。勿論、既に住基ネットを利用した全国民と在留外国人等への付番は済んでおり、行政や保険者等がマイナンバーの提示がされなかった際に、職権で地方公共団体情報システム機構(J-LIS)を参照し、マイナンバーを記入する事ができるとされている。この時、大元となる住民基本台帳に記載されていない情報は、当然ながら参照不可能である。その為、例えば類似する住所の誕生日もほぼ同じといった同じ漢字を用いた、読みが同じか又は読みの違う複数人や、性別情報に中途変更がされていたり・されていなかったりした場合、その判別は非常に困難であり、人為的不作為の誤謬を完全に抑止する事はそもそも無理である。そして、前述の漢字コード問題により、使用している端末の問題によって異なる文字が表示されたり、目視では（ほぼ）同じに見える文字が内部コードの問題で全くの別人として扱われたりする結果を招く事態は容易に想像可能である。何故ならば各市町村役場の端末内のデータを変換しながら全国ネットに繋げている為、文字コードを原因とするミスは、再点検でも同じプロセスを辿って再現され、見逃される可能性が高い。そもそも文字コードについては、ベンダーやプログラム、バグといった問題ではなく、データベース自身の仕様の瑕疵であることを認識しておかなければならない。

※行政のデジタル化には普遍、又は上位互換の保証されたコードの制定が先

現在マイナ保険証の誤登録問題について、政府の対応は後手後手、そして弥縫的であり、その原因を修正しているとは思えない。名前を表現する為の文字コードは勿論のこと、住所についても旧 JIS では付番されたコードが存在した。しかし、平成の市町村大合併等を経て、その変遷に追従した適切な更新や、その履歴は保全されているのだろうか？　住所の読みについても同じ文字で読みが異なると場所も異なる場合や、ｘｘ町とｘｘ本町といった、類似の地名が近郊に存在する場合、郵便番号の様

にコード化されていれば間違いは起こり難いが、文字情報のみの場合には信頼性は低下する。そして京都の地名表記の様に「通り」が基準で家屋に地番を振っていない地域や、「一丁目２番地３号」の様な表記とハイフンやマイナス記号で「１－２－３」とか「の」「ノ」で区切られた表記とかの同一性は整理されているのであろうか？

令和６年４月からマイナカードでの健康保険の資格確認が義務化とされた。そして、他人との紐付けや負担割合の間違いといった報道が後を絶たない。マイナンバー制度の適用範囲を拡げれば、その元となった住民基本台帳ネットワークの誤謬若しくはコード化の問題が解決されていない限り、問題事例が拡大するのは自明である。折角 IPA が文字情報基盤整備事業として文字情報基盤の文字（ＭＪ文字）を纏め上げたのであれば、先ずは住民基本台帳の文字コードの混乱を修正完了するべきであり、その結果としてマイナンバー制度に相当するシステムの正確性を上げ、それから他の行政情報等との紐付けを行うのが筋である。今の文字と文字コードの対応すら一意ではないシステムでは、どんなに再点検を行っても淘汰するのは不可能であろう。

※ＭＪ文字コードの更なる混乱

デジタル庁が、地方公共団体で利用される基幹業務のシステム統一・標準化を進める為に必要な要件の一つに漢字の統一がある。（2023 年時点でマイナ保険証の原則利用義務化が決定しているのだが）法務省は 2024 年３月を目途に戸籍ベンダーが管理している約 163 万文字の内、文字情報整備作業で文字情報基盤に同定できたもの約 55 万文字は 2017 年（基本は 2011 年）の文字情報基盤文字セット、通称ＭＪ文字（約６万文字）に収斂されているが、ＭＪ文字に同定できなかったもの約 15 万文字を約５万文字に収斂させたものを併せた約１１万文字を「戸籍副本システムにて連携する文字」として定めた。これから実際に戸籍に使用されている文字等を絞込む作業を行い約１万文字としたものをＭＪ文字に加え、戸籍の運用上必要な文字としてＭＪを拡張した文字セット(ＭＪ＋)約７万文字として全ての政府の標準準拠システム間において氏名等を情報連携する場合には、ＭＪ＋を利用すると定めた。尚、この文字要件対応スケジュールによれば標準化完了は 2025（R7）年度第４四半期末とされており、文字管理運用開始は 2026（R8）年度からとされている。（デジタル庁の文字要件説明資料を見ても、毎回内容や時期に度々の変更がされていて、その混迷ぶりが窺い知れる）

そもそも、データベースを運用するにあたって、運用開始後の文字コードの変更は有ってはならない事である。文字コードが収斂し少ない文字セットになるのならば、対応表を用いれば一方向の変換は一意的に可能だが、文字セットが肥大化すると、対応の維持は不可能となる。何故ならば、データの登録時期によって個人を特定する氏名の文字コードが変化すれば、当然別データとして扱われるからであり、その間違い防止の為にマイナンバーを付番した筈なのに、その元となる戸籍或いは旧住基ネットそのものを構成する文字が変化し続けていれば、その信頼性は揺らぐ。従って若し国民に付番

をするのならば、それが今のマイナンバーの様に目に触れる形であるかないかに関わらず、少なくとも文字コードの確定が先であるべきである。

※人為的ミスの誘発蓋然性は高い

政府の標準準拠システム間において約７万文字のＭＪ＋文字の利用をとの方針だが、そもそも戸籍其の他を扱う自治体職員をはじめ、健康保険組合等で入力を担当する職員にとって、７万文字の識別は誤謬無く可能なのであろうか？　年末「今年の漢字」としてニュースで清水寺で揮毫(きごう)される模様が流れる事でも知られる、俗に「漢検」として知られる公益財団法人・日本漢字能力検定協会が行っている日本漢字能力検定試験で、１級：大学・一般程度（約 6000 字）、準１級：大学・一般程度（約 3000 字）、２級：高校卒業・大学・一般程度（2136 字）（以下１０級まで省略）とされている。1978 年の旧 JIS C6226-1978（現 JIS X 0208 相当）では常用漢字 1945 字（当用漢字の後継として 1981 年（昭和 56 年）内閣告示された常用漢字表 1945 字とほぼ同じ）とその他の人名用漢字から第１水準漢字 2965 字が規定された。

これは平成２２年内閣告示の（現行の）常用漢字表の本表には 2136 字種が挙げられており、ほぼ漢検２級相当として合理性のある数であろう。翻って漢検の１級は非常に難易度が高く、合格率は毎回５～10%程度、さらに合格者の多くがリピーターで、新規合格者の割合はさらに低く、1～3%程度であり、具体的な人数としては、試験ごとに 50～100 人程度が合格しているとされている。即ち、データ入力担当者全員が漢検１級を持っていたとしても、約６０００字の識別が限界であり、その１０倍を超すＭＪ＋文字約７万字を扱う時点で、既にその要求能力において破綻していると言わざるを得ない。人為的ミスは起こるべくして起きると預言できる。竹槍でＢ２９が落とせないのと同じである。

（以上、未定稿）