Keywords you should know. 第117回

知っておきたいキーワード

文字符号

(正会員) 武 智 秀 †

† NHK エンジニアリングシステム
"Character Coding" by Masaru Takechi (Advanced R&D Department, NHK Engineering System Inc., Tokyo)
キーワード：文字符号，ISO-2022，UCS, UTF-8，異体字

そもそも文字符号とは何か？

コンピュータシステム上などで文字を扱うために，文字符号と呼ばれる符号が利用されています．文字符号はソフトウェア間や機器間での効率のよい文字情報の受け渡しや処理を可能にしています．Webページの記述や伝送，コンピュータプログラムの記述やコンパイラによる処理，ワードプロセッサによる文書処理など文字符号がなければ困難なことがたくさんあります．文字符号はその名の通り，文字を符号値で表すためのものですが，ではまず，そもそも符号値を割当てる対象である文字とは何なのか考えてみましょう．

図1の3文字はすべて図形が違いますが，この3文字の場合，図形が違うという理由で違う文字だと考える人はいないと思います．これらは同じ文字と考えるでしょう．つまり，文字とは，特定の図形で表されるものではなく，図形から「連想される形」(これを字体といいます) によって識別されているのです．字体は人間の頭の中にしか存在せず，紙やディスプレイに表示されるものはフォントや活字を使って「ある形にしてみた」もの (これを字形といいます) でしかありません．これは言語によらず同じです．頭の中の存在である字体それぞれに対して符号値を割当てたものが文字符号です．この関係を**図2**に示します．したがって，文字符号の規格書にはその規格で扱う文字の範囲を定めるためにたくさんの文字を掲載した表がありますが，表に挙げられている文字はすべて何らかのフォントを用いた字形で表現されているので「例示」なのです．このように文字符号とフォントは切り離されているので，**フォント**や***書体***の変更や，コンパイラやXMLパーサーなどによる文字列処理が非常に容易になるのです．まずは，文字符号と文字の図形やフォントは別のものであるということを覚えておいてください．

図1 同じ字？？

図2 字形と字体と文字符号

符号化文字集合と文字符号化方式

文字符号は，規格によって定められます．その際，その規格が対象とする文字の集合を定めます．これと符号値を明確に関係づけたものを符号化文字集合と呼びます．例えば，いわゆるJIS第4水準までの漢字を規定しているJIS X0213:2004という規格では，ひらがな，カタカナ，漢字，数字や記号など合計で11,233文字からなる文字集合を規定しており，それぞれの文字と符号値の関係を規定しています．これがJIS X0213:2004の符号化文字集合です．

符号化文字集合の符号値は，他の符号系との整合やデータ伝送時の安全性を高めるために符号化（エンコード）されることがあります．この方式を文字符号化方式と呼びます．JIS X0213:2004では，Shift_JIS-2004，ISO-2022-JP-2004，EUC-JIS-2004の符号化方式が記述されています．

ISO-2022とUCS

文字符号の規格には，いくつかの種類があります．日本語において現在利用されている文字符号体系は大別すると次の2種類があります．

・ISO-2022ベースの符号
・国際符号化文字集合（UCS）

ISO-2022ベースの符号は古くから用いられてきた符号系で，英語用に開発された1バイトの符号体系を拡張して漢字のような多数の文字や多言語を扱えるように考えられたものです．ISO-2022は，メッセージの送信者と受信者の間での取り決めに基づく独自の拡張を許容しているため，多くの派生規格があります．**図3**にその一つであるARIB STD-B24で規定する体系を示します．これは現在のデジタル放送で用いられている文字符号です．エスケープシーケンスによって符号化文字集合を切替え，8単位符号表に割当てることで多くの文字を扱えるようになっています．JIS X0213:2004で述べられているISO-2022-JP-2004，EUC-JIS-2004も，ISO-2022ベースの符号化方式です．

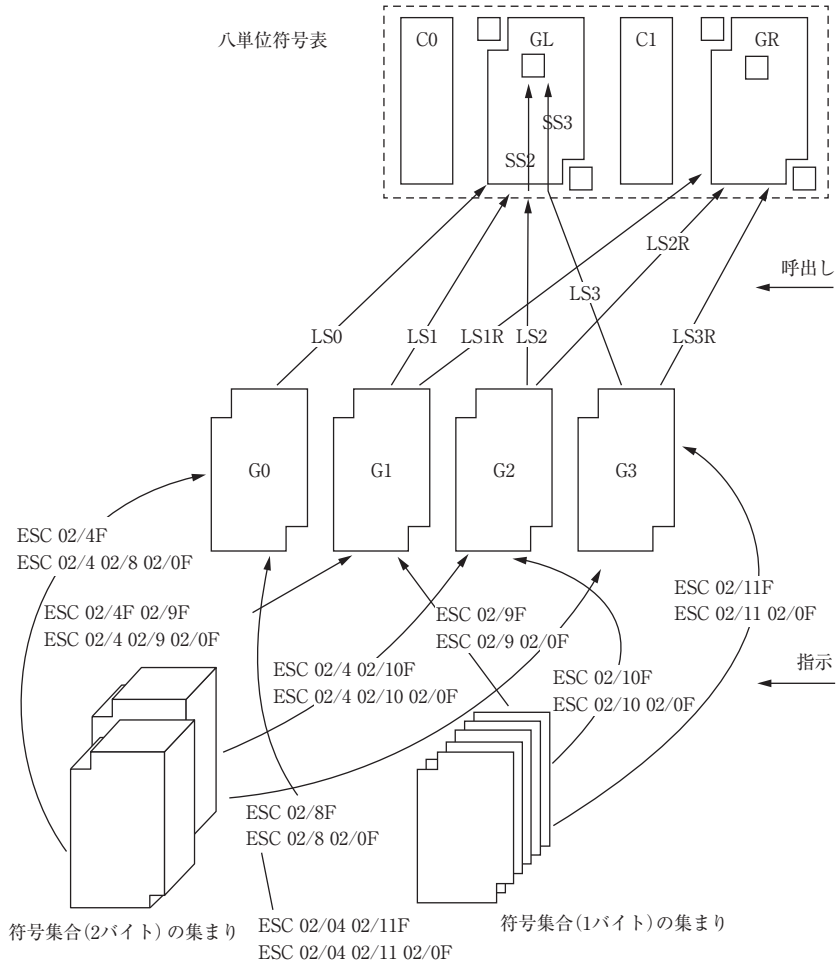

図3　ARIB STD-B24の文字符号の構造

これに対して，切替えではなく大きな符号空間であらゆる言語に対応しようとして開発されたのがUCS（Universal Coded Character Set）です．UCSはISO/IEC 10646およびUnicode Standardとして規定されており，もともと4バイト（32ビット）の空間が規定されていましたが，最新の規格では，先頭の約100万文字分（それぞれ2バイトの空間を持つ0x00面から0x10面までの17面）の領域を使うこととなっています．UCSの構造を**図4**に示します．

UCSには多くの現代の言語の文字や音楽記号，古代エジプトヒエログラフ，絵文字など多様な文字が収録されています．Unicode Standard 9.0.0では12万文字以上が規定されています．文字数が多い漢字については，日本語，中国語，韓国語，ベトナム語で用いられる漢字を統合したもの（CJK統合漢字と呼びます）として規定されており，日本のJIS規格で定められている ☞

☞ 漢字も収録されています．CJK統合漢字はJIS規格を大きく超える87,000以上の文字が規定されています．

UCSでは2バイトの空間である面を17面使うので21ビットの符号化空間で考える必要があります．一番手前の面(符号値でU+0000-U+FFFF)を基本多言語面（Basic Multilingual Plane: BMP）といい，それ以外の面にもそれぞれ名前があります．JIS X0213:2004の漢字はBMPと第2面である補助漢字面（Supplementary Ideographic Plane: SIP，U+20000-U+2FFFF）に配置されています．

UCSはPC，タブレット，スマートホンなど多様なデバイス，ソフトウェアの上で広範囲に利用されており，現在，主流の文字符号となっています．UHDTV放送用の規格であるARIB STD-B62もUCSに基づいた文字符号を規定しています．

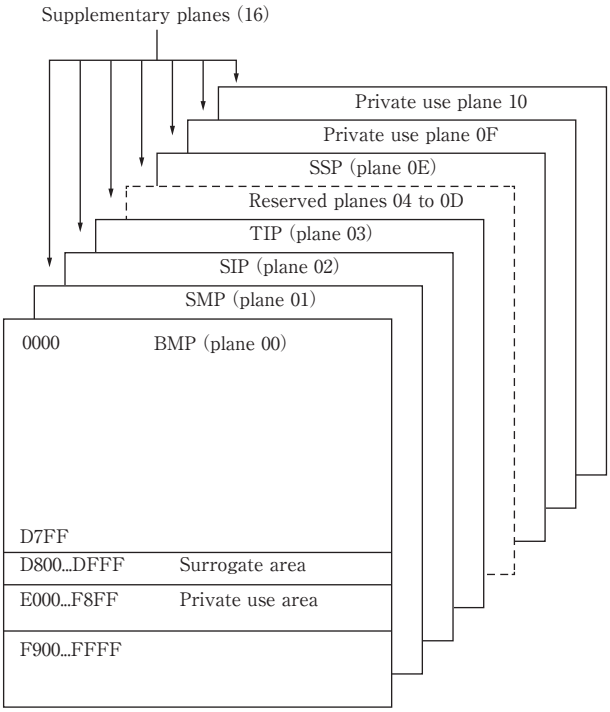

図4 UCSの構造

UTF-8

UTF-8はUCSの文字符号化方式の一つで，XMLや一部のプログラム言語において必須とされています．またUCSにおける文字符号化方式として最もよく使われているものです．

UTF-8の符号化法を**図5**に示します．U+0000からU+007Fの範囲にはアルファベット，数字，記号が配置されていますが，この配置はASCIIなどの他の文字符号と基本的に同じです．この範囲の文字については最下位1バイトだけを伝送することで，UCS以前に開発された処理系との互換性を高く保っています．一方，UCSの広大な符号化空間を扱うために，UTF-8は1バイトから4バイトの可変長符号となっており，ISO-2022ベースの符号のようにバイト数からは文字列の長さは判断できません．

図5を見ると，2バイト以上のエンコード値になるものでは第2バイト以降は必ず“10”で始まり，第1バイトが“10”で始まることがないので，第1バイトか否かの見分けがつきます．第1バイトを見ると，最上位ビット（MSB）からの“1”となっているビット数がその文字のエンコード値のバイト数を表しています．また，MSBが“0”であれば，U+0000-U+007Fの範囲の文字であると判断できます．つまり，UTF-8を用いると，可変長符号であってもバイト列を検査すると各文字の先頭の位置がわかるのです．簡単には，UTF-8のバイト列に対して，0xC0とのANDを取るとその結果で先頭のバイトか否かがわかります．0xC0であれば2バイト以上のエンコード値の第1バイトであり，0x80であれば2バイト以上のエンコード値の第2バイト以降，それ以外であれば1バイトのエンコード値になります（実用的な先頭の判定のためには，誤り状態の判定のためにもう少し☞

UCS符号値	ビットパターン	第1バイト	第2バイト	第3バイト	第4バイト
U+0000-U+007F	000000000xxxxxxx	0xxxxxxx			
U+0080-U+07FF	00000yyyyyxxxxxx	110yyyyy	10xxxxxx		
U+0800-U+FFFF	zzzzyyyyyyxxxxxx	1110zzzz	10yyyyyy	10xxxxxx	
U+10000-U+10FFFF	000uuuuuzzzzyyyyyyxxxxxx	11110uuu	10uuzzzz	10yyyyyy	10xxxxxx

図5 UTF-8の符号化方式

☞ 複雑な手順が必要です）．「あいAB剱」というUTF-8でエンコードした様子を**図6**に示します．この性質を使えば，データ伝送時に誤りがあったとしても，誤りを伝播させずに誤ったデータを含む文字だけに限定させることが可能です．この性質はUCSの他の文字符号化方式やISO-2022にはないものです．特にISO-2022ベースの文字符号では，こうした性質を規格として持たせることが困難です．ISO-2022ベースの文字符号では，図3に示すとおり，多くの符号化文字集合をエスケープシーケンスによって切替えるため，復号する側において，現在対象としている符号化文字集合が何であるかの「状態」を把握している必要があります．一旦状態が把握できなくなったら，次に状態を設定するエスケープシーケンスが来るまでは正しい復号を行うことができないので，UTF-8は，伝送誤りに対して大きなアドバンテージがあると言えます．

「あいAB剱」をUTF-8で符号化すると・・・・

	あ			い			A		B		剱		
UCS符号値	30		42	30		44	00	41	00	42	2	07	4F
UTF-8エンコード値	E3	81	82	E3	81	84	41		42		F0	A0	9D	8F
0xC0とAND	C0	80	80	C0	80	80	40		40		C0	80	80	80

各文字の先頭バイト

図6　UTF-8による符号化の様子

包摂と異体字

図1の文字は，三つとも同じ文字でした．では**図7**の文字はどうでしょう？人や場合によって意見が分かれるのではないでしょうか？

図7の(a1)と(a2)，(b1)と(b2)は異体字と呼ばれる関係です（図7は有名な例で，「クチだか」と「ハシゴだか」，「シよし」と「ツチよし」と呼ばれます）．異体字は，字の意味や音が同じ同一の字種ではあるものの字体が異なっている，というものです．一般には異体字は相互に入れ換えても同じ文脈として成立します．図7の(a1)と(a2)，(b1)と(b2)はそのように考えられる文字の組の例です．異体字は字形のゆれが大きい漢字に顕著ですが，モンゴル文字などにもあります．

JIS X0213の目的の一つに，情報の交換性を確保することがあります．そのため，異体字のようなものの場合，複数の字体を区別せずに同じ符号値に対応させます．これを包摂といいます．

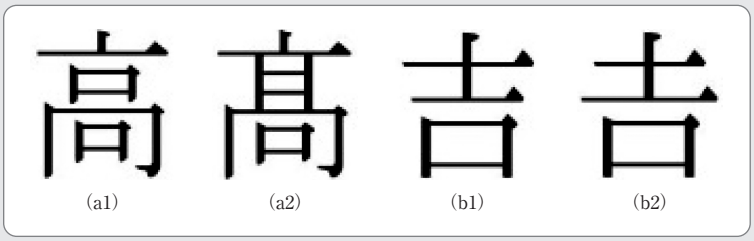

図7　異体字の例

包摂を行うにあたって，JIS X0213では詳細な包摂基準を設けています．ある文字が別の文字と同じ文字なのか異なる文字なのかを具体的に示す基準を設けたのです．包摂基準は過去の膨大な文献を調査した上で約200の基準が定められています．つまり，JIS X0213で包摂されている文字には，それぞれその根拠があるのです．

JIS X0213では，包摂基準によって図7の(a1)と(a2)，(b1)と(b2)は同一の文字であるとされています．したがって，(a1)と(a2)，(b1)と(b2)は同じ符号値となります．フォントの実装においては，具体的な図形文字の字形としては，(a1)でも(a2)でも，また(b1)でも(b2)でもどちらでもよいということです．逆に(a1)と(a2)は同じ文字ですから，例えば，外字として異なる符号値を割当てて(a1)と(a2)を識別することは重複符号化となり，この規格の立場とは相容れないことになります．

一方で，図7の(a1)と(a2)，(b1)と(b2)を識別する必要がある，というニーズもまた存在します．例えば，電子政府では人名や地名をコンピュータの上で扱いますが，法律の定めによって異体字を識別しなくてはならない場合などがそれにあたります．

異体字タグ

異体字の問題に対処するため，UCSでは異体字タグ（Variation Tag）というものが定められています．このうち，漢字に関するものはIdeographic Variation Selector（IVS）と呼ばれ，第14面（Supplementary Special-purpose Plane: SSP，U+E0000-U+EFFFF）に配置されています．

文字符号は「連想される形である字体」に対して符号値を割当てるものですが，IVSに関しては「具体的な図形」に対してセレクタ値を割当てます．IVSを用いると，**図8**の（c1）は<U+8FBB:U+E0103>，（c2）は<U+8FBB:U+E0102>として表されます．この「U+E0103」，「U+E0102」がIVSです．

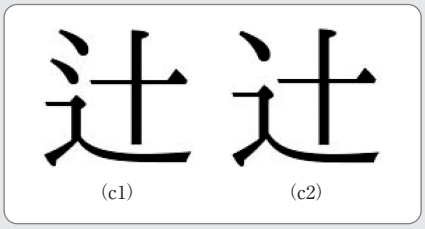

図8　IVSで識別される字の例

IVSを使うと，正しく実装された処理系においては，U+8FBBで検索すると（c1），（c2）どちらもヒットします．IVS込みで検索すると，（c1）または（c2）どちらかのみをヒットさせることもできます．

IVSは，データベース化されている異体字の集合のコレクションに対して割当てられています．このデータベースをIdeographic Variation Database（IVD）と呼びます．漢字に関しては，Adobe-Japan1，Hanyo-Denshi，Moji_johoの3種があり，それぞれUnicode技術委員会のWebサイトで公開されています．

IVSは便利なメカニズムですが，まだ発展途上です．また，CJK統合漢字の中には，CJK統合漢字の源規格分離則（Source Separation Rule）によって異なる符号値を持つものもあります（図7の二つの例は，これによってIVSではなく（a1），（a2），（b1），（b2）それぞれ異なる符号値を持っています）．しかし，新たな漢字の追加提案に対して，IVSを積極的に活用した標準化の推進も提唱されています．

むすび

文字符号は文化と技術の接点で，非常に奥が深い符号です．また，最も基本的な情報表現の一つであり，一旦定めたらそれを用いてデータが蓄積されるため，二度と方式や文字の変更・削除はできません．一方で小説，記録，文書，コンピュータプログラムなどさまざまな文化的産物がこれを利用するため，それぞれのニーズを満たしつつ現実的な方法として電子的に扱うために，現在でも休むことなく開発が続けられています．規格書そのものは難解ですが，さまざまな本やWebサイトで解説が多くあります．文字化けを嘆くだけでなく，その理由に思いをはせれば，いろいろなことが見えてくると思います．

（2017年6月21日受付）

武智（たけち）　秀（まさる）
1990年，東北大学大学院工学研究科（電子工学専攻）修士課程修了．同年，NHK入局．以来，同放送技術研究所にて，衛星放送システム，ディジタル伝送方式，マルチメディア方式，ディジタル受信機アーキテクチャなどの研究に従事．また，デジタル放送およびそのマルチメディア方式について，ARIB，IPTVフォーラム，ITU-R，ITU-T，ISO/IECにおける国内，国際標準化に参画，ARIB，ITU-R，ITU-Tでの作業班主任やSWG議長，ラポータ等を歴任．現在，ITU IRG-IBB共同議長，（一財）NHKエンジニアリングシステム先端開発研究部上級研究員．正会員．

キーワード募集中

この企画で解説して欲しいキーワードを会員の皆様から募集します．ホームページ（http://www.ite.or.jp）の会員の声より入力可能です．また電子メール（ite@ite.or.jp），FAX（03-3432-4675）等でも受け付けますので，是非，編集部までお寄せください．

（編集委員会）