解説

世界の文字と文字符号
（後編）

三上 喜貴
長岡技術科学大学
mikami@kjs.nagaokaut.ac.jp

ザバルスキー パヴォル
長岡技術科学大学
zavarsky@vos.nagaokaut.ac.jp

前編では世界の言語と文字を概観するとともに，文字体系を 5 つのカテゴリーに分類し，そのうち，アルファベットと単子音文字について述べた．後編では，インド系文字を含む音節文字と漢字を扱い，最後にこれまでの文字符号化の到達点ともいうべき国際符号化文字集合 ISO/IEC 10646 の意義と利用の現状，残る課題について述べる．あわせて筆者らが進めている言語天文台の活動について紹介したい．

○ 表意文字由来の音節文字 ○

音節文字（syllabics あるいは syllabary）とは，「音節を単位として表記される表音文字」であり，音節文字の総数はその言語の表現に必要な異なる音節の総数で決まる．では音節文字を用いる諸言語の音節総数はいくつなのか．

日本語の音節数を考えるには森田式キーボードが参考になる [1]．この方式は日本語の音韻構造の特性を巧みに利用したもので，音節（CV[C]）を声母（C-）と韻母（-V[C]）に分ける中国語の「半切」と同様の分析を行って，和語・漢語の入力に必要なキーの種類を決定した．結果のみを示せば，声母に相当するものの種類が 30，韻母に相当するものの種類が 30，声調はないから音節総数は最大で 900 ということになる．しかし，日本語の仮名は音節文字といっても，促音，拗音などを考えれば 1 つの音節を常に 1 つの文字で書くわけではないから実際にははるかに少ない仮名文字集合で用が足りているのである．

中国の四川省，雲南省などに居住する少数民族である彝（い）族が使用している彝文字は元来表意文字であったが，これから派生した音節文字である現代彝文字は音節と文字が完全に 1 対 1 で対応し，しかも声調の異なる音節にも別の文字を用意しているから文字集合の総数はかなり多い．中国政府はこの少数民族文字のために国家規格 GB 13134:1991 信息交換用彝文編碼字符集を制定し，さらにこの文字集合はほぼそのまま ISO/IEC 10646 に継承されて合計 1,149 字の符号位置が割かれることとなった．

仮名と彝文字は共に表意文字を原型として形成されたから，同一の子音や母音を含む音節といえども，その字形の間にはまったく共通性が見当たらない（**表 -1**）．したがってこれを図形的な要素に分解することは不可能ないし困難であり，そもそも分解しようという発想すら生まれない（もっとも手旗信号はカナを図形的要素に分解して符号化したものともいえる）．符号化にあたっても，音節文字のそれぞれに独立した符号を与えるのが自然な流れであって，音節文字の符号化にはあまり混乱要素はない．

○ 結合音節文字―インド系文字とハングル ○

これに対して，インド系文字やハングルは子音と母音を表す図形要素を組み合わせて表現される音節文字であるために，同一子音や同一母音を含む文字は外観上も明

ひらがな	-a	-i	-u	-e	-o
k-	か	き	く	け	こ
s-	さ	し	す	せ	そ
t-	た	ち	つ	て	と

彝文字	-i	-ie	-a	-uo
b-	ꀀ/ꀁ/ꀂ	ꀃ/ꀄ/ꀅ	ꀆ/ꀇ/ꀈ	—/ꀉ/ꀊ
p-	ꀋ/ꀌ/ꀍ	—/ꀎ/ꀏ	ꀐ/ꀑ/ꀒ	—/ꀓ/ꀔ
m-	ꀕ/ꀖ/ꀗ	—/ꀘ/ꀙ	ꀛ/ꀜ/ꀝ	ꀞ/ꀟ/ꀠ

タミル	-a	-ā	-i	-ī	-u	-ū	-e	-ē
k-	க	கா	கி	கீ	கு	கூ	கெ	கே
nn-	ண	ணா	ணி	ணீ	ணு	ணூ	ணெ	ணே
r-	ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரே

デーヴァナーガリ	-a	-ā	-i	-ī	-u	-ū	-ri	-ē
k-	क	का	कि	की	कु	कू	कृ	के
r-	र	रा	रि	री	रु	रू	—	रे
h-	ह	हा	हि	ही	हु	हू	हृ	हे

ハングル音節文字	-a	-i	-u	-e	-o
k-	가	기	구	게	고
n-	나	니	누	네	노
t-	다	디	두	데	도

注）彝文字には 3 種類の声調があり，声調によって字形が異なる．彝文字の各欄に "/" で区切って示した 3 種類の文字は，これらの各声調に対応した文字である．□で囲った文字は規則性が破られて変則的な合字が形成されているケースである．インド系文字ではしばしばこうした合字が見られる．

表-1　さまざまな音節文字

らかな共通性がある（表 -1）．インド系文字の場合には，基本となる子音文字の総数はたかだか 30 前後とアルファベットに近く，また，基本となる子音文字にさまざまな記号が付加されて母音の変更が行われる様子は見かけ上はアルファベットに補助記号が付加される様子に似ている．このためインド系文字をアルファベットと音節文字の中間という意味で alphasyllabary と呼ぶ者もいる．こうした特長を有するために，結合音節文字の符号化にあたってはいくつかの選択肢が生まれ，実際にさまざまな符号化方式が各国で考案されてきた．(1) 音節文字をそのまま符号化する方式，(2) 図形的要素に分解して符号化する方式，(3) 音韻的要素に分解して符号化する方式の 3 つである．

音節文字をそのまま符号化する方式はいわば活字方式である．インドの印刷史研究者 Priolkar は，ゴアで活動を行っていたあるイエズス会士がローマに宛てて書いた 1608 年の書簡の中に，『私は永年にわたってこの地（マラバル地方）の言葉と文字で書籍を印刷しようと努力してきましたがまだ実現できていません．その第 1 の理由は，ヨーロッパではわずか 24 種類の活字を準備すればよいのに対して，この地方の文字は音節単位の文字であり，その活字を鋳造するためには 600 を超える鋳型を作らなければならないという困難があることです』という記述を発見している [2]．最近に至るまで，インドの活字印刷所の多くは，言語によって差はあるもののおおむね同規模の活字を揃える必要があった．しかしながら，コンピュータ用の文字符号としてインドでこのような大規模な文字符号表が作られることはなく，後述するように（2）と（3）の方法が採られた．

一方，韓国語のハングル文字も CV[C] 型の音節文字であるが，韓国の国家規格 KS C 5601: 1987 Korean graphic character set for information interchange という 2 バイト符号表が作られ，ハングル文字に 2,350 字を割り当てたのは，漢字文化圏に位置したためであろう（ISO/IEC 10646 では論理的に可能な組合せである 11,172 音節のすべてに符号を割り当てた）．インド系文字と同様の特徴を持つエチオピア文字の場合，音節数はハングルやインド系文字ほど大きくはないものの，ISO/IEC 10646 においては，やはり音節単位の符号化方式が採用されている．

音節文字を図形的要素に分解して符号化するという方式は機械式タイプライタに起源を持つ．**図 -1** に示した 8 台の機械式タイプライタのうち，英文タイプを除く 7 台はいずれもアジアの言語にローカライズされたタイプライタである．これらの言語の使用する文字はまったく異なるにもかかわらず，タイプライタの外観，鍵盤の総数はほとんど同じである．実際のところ，これらの機械の製造者はオリンピア（米），アドラー（独）など限られた欧米メーカであり，基本となる機構は共通である．タイプライタにおける鍵盤数の制約は絶対的であり，タイピングアームの先端に上下段あわせてたかだか 90 の異なるタイプフェースを貼り付けることができるに過ぎない．残された工夫の余地は，印字すべき文字集合をなるべく少数の図形要素に分解し，その重ね打ちによって

図-1　筆者のタイプライタ・コレクション

上段左からタミル語，ベンガル語，シンハラ語．中段左から英語，ベトナム語，韓国語．下段左からミャンマー語，タイ語である．韓国語タイプライタは，初声（頭子音），中声（母音），終声（末子音）の重ね打ちによってハングルを印字している．

もとの文字集合を再現するというパズルを解くことである．そして，これらのタイプライタ開発者たちは試行錯誤の末に，さまざまなバリエーションの鍵盤配列を考案した．コンピュータの時代に入って，タイプライタ鍵盤上の図形は，今度は符号表上に移し替えられた．ASCII文字集合はおおむねタイプライタの文字集合を継承しているから，どんな鍵盤配列をモデルにするにせよ，この符号化方式を採用することの利点は英語版ソフトの上での実現が容易なことである．この結果，簡便なローカライズ方法として多くのローカルベンダが採用するところとなったが，同時に，タイプライタ時代の鍵盤配列の混乱がそのままコンピュータ上の文字符号にも継承されるというマイナス効果ももたらすこととなった．フォント開発者には符号表を作っているという意識は希薄であり，ある者はこうして作られた符号を accidental internal code と呼んだ．筆者らもヒンディー語だけでも 20 種類，タミル語については 17 種類のこうした符号化方式を確認しているが，実際にはさらに多数が存在しているであろう．タイ文字符号に関する国家規格 TIS 620:2533（西暦年では 1990）Thai character codes for computers はこの方式を採用した符号表であるが，1990 年という比較的早い時期に標準が成立した 1 つの理由は，安定したタイ語タイプライタの鍵盤配列標準を持っていたことである．

最後に，音節文字を音韻的に分解して符号化する方式について述べよう．ISO/IEC 10646 におけるインド系文字符号の原型となったインドの国家規格 IS 13194:1991 Indian script code for information interchange（ISCII）は，インド各州公用語で使われている 10 種類の文字を，共通の枠組みによって音韻的に分解することによりわずか 7 ビットのコード表で表現する道を開拓した．いわば，音節文字を仮想的なアルファベット（単音文字）の符号列によって再構成する方式である．インド系文字の基本子音，基本母音の数は合計してもたかだか 60 前後であり，大文字，小文字をあわせたアルファベット文字集合と大差ない．したがって，この方式もまた符号表をコンパクトに抑えることができ，まことに論理的な方法であるが，その半面，符号列から表示文字を合成するレンダリング機能は著しく複雑なものとなる．専用ソフトウェアを必要とするため，IS 13194 がインターネット上で広く使われることはなかった．

このように見てくると，文字符号の開発において最も混乱が著しいのがインド系文字であることが分かろう．国際的に認知されているエスケープシーケンスの登録簿[3] やインターネット上で利用される文字符号化方式の登録簿である IANA リスト[4] を見ても，インド系文字に関しては，唯一タイ文字規格 TIS 620 が登録されているだけである．ISO/IEC 10646 の誕生によって，これらの混乱の収束に向けた道が用意されたかに見えるが，まだその実際の利用が広がるに至っていないことは後ほど紹介するとおりである．

○ 漢字 ○

表意文字の文字集合総数は，その字義通りに考えれば語彙の総数と同じである．17 世紀に中国で布教を始めた Mateo Ricci 神父は，「中国には単語と同じだけの文字がある」と述べて驚きを示したという．漢字が表語文字として発達したのは，中国語が非屈折語☆1 であって，非膠着語☆2 であるという条件が大きく作用した．名詞に格変化があり，動詞に時制の変化があるとき，表意文字を用いてこれを表現するためには何らかの工夫がいる．膠着語である日本語は漢字を取り入れるにあたって「テニヲハ」を表記するために具体的な語彙と結びつかない表音文字体系を考案する必要があったし，ハングルを考案した韓国語にも同様の必要性があった．

符号化という観点から見るとき，巨大な表意文字集合である漢字については，これを何らかの要素に分解して符号化するという選択は考えにくいから，必要に応じたサイズの符号表を用意するしかない．こうした巨大文字集合を簡便に利用できるかどうかは，結局のところ入力

☆1　屈折語とは，語の文中における文法的な役割や関係の差異を，語形の一部を変えて表す言語で，主として語尾変化として現れる（大辞林）．

☆2　膠着語とは，実質的な意味を持つ単語あるいは語幹に，文法的な機能を持つ要素を次々と結合することによって文法的な役割や関係の差異を示す言語で，朝鮮語，トルコ語，日本語，フィンランド語などが該当（同上）．

地域	1960 年代	1970 年代	1980 年代	1990 年代～
ラテン文字圏	ASCII/ISO 646	欧州言語への拡張（6937, 8859）		
キリル文字圏	GOST 13052	露語以外の諸語への拡張		
アラビア文字圏			ASMO 449	
ヘブライ文字圏			ECMA 121	
日本	JIS C 6220	JIS C 6226		
中国			GB 2312	少数民族文字
韓国			KS C 5601	KS X 1005
タイ			TIS 620	
インド			ISSCII 83	IS 13194
ベトナム				TCVN 5412
スリランカ				SLS 1134
国際符号化文字集合				ISO/IEC 10646

図-2　各言語圏における文字コード形成の歩み

方法の問題に帰着する．日本で，JIS 漢字コードの成立（1978 年）と同時にカナ漢字変換方式による第 1 号ワープロが誕生して日本語情報処理時代の幕を開けた事実はこの関係を物語っている．2 年遅れて，中国でも GB 2312:1980 信息交換用漢字編碼字符集基本集（収録漢字 6,763 字）が成立した．

音節文字の場合に符号化方式に関して図形的分解，音韻的分解といった方策が模索されたのと相似的に，漢字においては，図形的分解，音韻的分解といった方策が入力段階における選択肢として現れた．遡れば，これは字典編纂者の頭を悩ませてきた配列順序や検字法と同根の問題である．図形的分解を基礎とする方式としては，部首画数索引，四角号碼☆3 などが，音韻的分解を基礎とする方式としては，韻書と呼ばれる方法などさまざまな配列･検字方法が字典編纂の歴史上開拓されてきた．これらは漢字の符号化方式の先駆をなすものであり，実際，情報交換用の漢字符号開発の初期には，こうした検字方式と対応する，入力方法と直結した符号化方式が多数考案された．筆者の手元にはこうした各種漢字符号の対応辞書ともいうべき『常用漢字編碼字典』[6] があるが，ここには，GB 2312，電報碼，大衆碼，三声碼，中文声数編碼，前三末一漢字輸入編碼，部形編碼，筆形編碼など，合計 23 種類もの符号体系の変換表が収録されている．

漢字の符号化をめぐる一連の問題群として，収録範囲や漢字の同定に関する問題，CJK の漢字統合に関する問題があるが，これについてはすでに多くのことが書かれているので本稿では省略する．

○ 国際符号化文字集合（UCS）○

以上，駆け足で世界の文字とその符号化の歩みをたどってきた．これを概観したものが**図 -2** である．そして，こうした文字符号開発の現時点における到達点として国際符号化文字集合 ISO/IEC 10646 Universal multiple-octet coded character set（以下，本章以降では単に UCS と略記）がある．これは 4 オクテット（正確には 2^{31}）という巨大な符号空間を用いて，世界中の利用者が必要とするすべての文字について，符号のみによって一意に文字を特定しようとするものである．芝野耕司はその意義を「①情報処理，通信，図書館の 3 分野における文字コードの統合，②情報の処理，交換，蓄積，入出力の全領域で共通に使用できる文字コードの開発，③二者間での合意に基づく情報交換から，より普遍的な情報交換を可能とする文字コードの開発」という 3 点にあると整理した [7]．UCS 登場以前の文字符号が何らかの意味で情報交換当事者間の取り決めを前提とした情報交換であるのに対して，UCS は，当事者間の事前合意を前提とせず，適用分野を問わず，世界の文字を符号（および文字名）のみによって一意に同定するものであるという意味において，文字通り普遍的な（"universal"）性格を持つ文字符号といえる．

では，UCS の普遍性を最大限に生かすべき Web ページの表記において，実際にはどの程度使用されているのか．筆者らの進めている言語天文台プロジェクト [5] の調査結果から 1 つの手がかりを提供しよう．言語天文台プロジェクトは，世界の Web ページを収集し，各ページの言語属性を使用言語，使用文字体系，使用文字符号の各側面から統計的に明らかにすることを目的としている．現在公開されている世界の Web ページ総数は 100 億ページを超えているものと思われ，そのデータ量はテキストに限っても数十テラバイトに達する．筆者らはまだアジアおよびアフリカ地域のドメインの一部を収集したに過ぎず，また言語属性判定ツールもまだ開発途上にあるので，本稿ではヘッダーに記載された charset 属性

☆3　漢字の四隅の形態的特長をもとに符号化する方法．

国・地域	ccTLD	国名	全頁数	UTF-8 頁数	使用比率
イスラム諸国会議機構(OIC) 加盟 57 カ国中の使用比率上位 10 カ国	gm	Gambia	192,860	161,790	83.89%
	af	Afghanistan	273,145	209,275	76.62%
	tm	Turkmenistan	219,357	165,974	75.66%
	ug	Uganda	241,972	159,574	65.95%
	ly	Lybia	322,659	157,904	48.94%
	ir	Iran	912,549	434,020	47.56%
	bd	Bangladesh	115,245	47,563	41.27%
	jo	Jordan	369,697	139,690	37.78%
	dj	Djibouti	406,145	141,969	34.96%
	al	Albania	174,313	55,376	31.77%
インド	in	India	1,382,909	87,972	6.36%
東南アジアのインド系文字使用国 4 カ国	kh	Cambodia	18,432	256	1.38%
	la	Laos	103,336	13,760	13.31%
	mm	Myanmar	27,213	5	0.01%
	th	Thailand	5,934,147	151,467	2.55%
ベトナム	vn	Vietnam	1,331,738	845,338	72.58%

表-2　Web ページのヘッダー記述からみたいくつかのドメインの UTF-8 使用比率

出典）言語天文台プロジェクト．OIC 諸国は 2004 年 11 月，インドと東南アジア諸国は 2004 年 7 月に取得したデータに基づいて算出した．

に基づく調査結果[☆4]のみを紹介する（**表 -2**）．

UCS の符号化方式としては UTF-8（8-bit UCS Transformation Format）が一般的である．これは UCS を 4 オクテットのままで符号化せず，最も使用頻度の高い ASCII は 1 バイトで，その他のアルファベット（本稿でいう単子音文字も含む）は 2 バイトで，それ以外の音節文字や漢字（ただし基本多言語面にあるもののみ）は 3 バイトで，という具合にして 6 バイトまでですべての文字を表現するという符号化方式である．いわば，ラテン文字に最適化された可変長符号であり，ラテン文字について言えば ISO/IEC 646 を使用しているのと変わらない．そこで，UCS のもたらす福音が本来最も期待される地域として，アラビア文字，キリル文字を含めた非ラテン文字利用言語を多数含むイスラム諸国会議機構諸国（合計 57 カ国，OIC: Organization of Islamic Conferences），インドおよび東南アジア 5 カ国（タイ，ラオス，カンボジア，ミャンマー，ベトナム）を取り上げてみよう．

まず OIC 地域を全体として見るとき，UTF-8 の使用比率は依然として 7% 強に過ぎないが，ccTLD（country code Top Level Domain）別に見ると UTF-8 使用比率が 70% を超えているドメインがいくつかある．トップのガンビアは英語が公用語だが，隣国のセネガル等でも使われているマンジェンゴ語は特殊なラテン追加文字（UCS ではラテン拡張 B に収録）を必要としており，2 位のアフガニスタンはパシュトゥー語，ダリ語などが拡張アラビア文字を使用する．いずれも UCS の登場によって安定的な符号表現が可能となった．東南アジアではベトナムの使用比率が最も高い．前編で述べたように，ベトナム語の表記に用いられるクオックグーは通常の ASCII 文字集合に対して 134 文字の追加を必要とする．このため，制御文字領域を使用するという「禁じ手」まで動員して国家規格 TCVN 5412 が制定され，このほかにも多数の符号化方式が乱立してラテン文字符号化方式の博覧会ともいうべき混沌状態を招いたが，ベトナム政府は 2001 年に至って UCS を国家規格 TCVN 6909 16-bit coded Vietnamese character set として制定し，その利用を促進する政策に転じた，という経緯を考慮すればこの結果は納得できるであろう．

総じて言えば，ラテン文字，キリル文字，アラビア文字をベースとした拡張文字集合の利用地域においては UCS の恩恵がもたらされつつあることを実感できるのに対して，インド，カンボジア，ミャンマー，ラオスなど，インド系文字圏における UCS 利用は依然としてきわめて低い水準にとどまっていることが分かる（タイの UCS 利用比率が低いのは現行のタイ文字符号 TIS 620 がすでに定着しているためであろう）．

○ 残された課題：インド系文字 ○

UCS にインド，ラオス，クメール，ミャンマー，スリランカなどの文字符号パートが登場して相当の年月が

☆4　ヘッダーに charset=UTF-8 と記述されているからといって実際に UTF-8 が使用されているとはいえない．逆に実際に UTF-8 が使用されている場合にはある程度の確かさをもってヘッダーにもそのように記述されていると推測できるから，ここで紹介する数値は実際の使用比率の上限値と考えていただきたい．

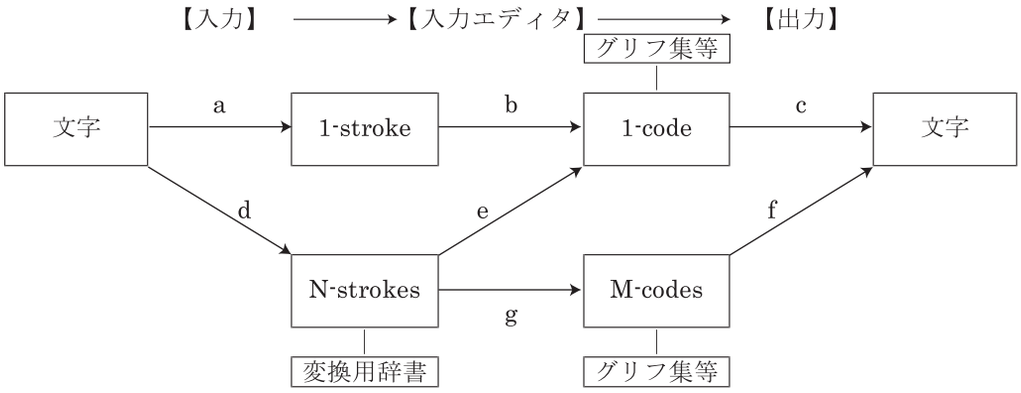

図-3　入力，入力エディタ，出力処理の流れ

経過しようとしているにもかかわらずいまだに利用が進んでいないのはなぜなのか．一般論として，小規模なユーザしか期待できない言語に対して膨大なコストをかけてローカライズすることは引き合わない商売であろう．あるシステムを1つの言語に対してローカライズするには，フォント開発，辞書開発，入出力エンジン，マニュアル作成，テストなどの費用として最低数億円が必要という．顧客が数十万人以上いなければ回収できない投資である．現実には違法コピー等の問題も加わり，ベンダから見ての投資意欲はさらに低下する．しかしこうしたベンダの行動ばかりが問題なのではない．UCS自身にも残された課題があるのではないか．

問題を整理するため，入力 → 符号表現 → 出力という一連の流れを**図-3**のように整理してみる．最もシンプルな英文のテキスト入力の場合，入力キーストローク，文字符号，出力に至る流れはすべて1対1で結ばれている（a→b→c）．アラビア文字のように複数表示形を持つ場合には出力時点で文脈を解析しながらグリフ選択を行う必要があるが，処理の流れはやはりアルファベットと同じである．日本語や中国語の場合，流れはd→e→cとなり，入力時点での入力エディタの負担は大きいが，出力時点は単純である．

これに対して，インド系文字のような音節文字の場合の流れはd→g→fである．入力すべき文字列は，まず入力者の頭の中でいくつかのキーストロークに分解され（d），何らかの処理を経て複数個の符号列として記録される（g）．検索，編集やソートはこの符号列を対象として行われる．そして，表示出力に際しては，レンダリングソフトが符号列を解釈しながら適切なグリフ列へと変換する（f）．このとき，符号列とグリフ列との関係はN対Mであり，この対応関係に関する言語固有の知識はレンダリングソフトが参照するグリフ集合や知識データベースにより与えられる．Windowsの場合にはOTF（Open Type Font）データベースとして，また多言語対応オープンソースソフトの場合には，それぞれの文字属性ファイルなどに格納される．表-2においてバングラデシュのUCS利用比率が41％にも達しているが，UCS利用ページを調べると，実は同国のLINUXグループの開設するページが大部分を占めていることが分かる．このように，今後，オープンソース開発者によってUCSの利用環境整備が牽引されるケースも増加しよう．

ここでISO/IEC 6937のことを思い出していただきたい．この規格は補助記号付き文字からなるグリフ集合をレパートリという形でも規定したが，インド系文字の場合にも，グリフ集合をトランスペアレントな形で規定することが必要ではないか．ただし6937の場合には合成の許される文字の種類を制限するのが狙いであったが，この場合には出力すべき達成目標をすべての開発者に対して明示するという狙いからである．UCSの符号化の際に，こうした巨大な音節文字集合のすべての要素に独立した符号位置を割り当てるという選択も論理的には可能であったはずだが，実際にはそのような選択は一部の文字（たとえばハングル，エチオピア文字，彝文字など）についてしか行われなかった．表音文字である音節文字を音韻的分解によってコード化するということにはそれなりの合理性があるから，このこと自体は否定的に捉える必要はないが，どこまで表示できる能力が必要なのかを示すグリフ集合の全体像および個々のグリフをどのような符号列によって表現するのかという対応規則に関して，実装者の裁量に任せられている現状には問題がある．これらの情報について，ベンダやオープンソースソフト開発者が共有することのできる何らかの仕組みを作り出すことが求められているではないだろうか，と筆者は考える．

○ 残された課題：歴史上の文字と少数民族文字 ○

現時点で，少なくとも各国の公用語で利用される文

字に関する限り文字符号開発の課題は終了したといってよい．先述したように，開発された文字符号の利用は必ずしも満足のいくテンポで進展しているわけではないが，新規の開発課題として残された対象は歴史上の文字と少数民族文字だけとなった．UCS の最新版である ISO/IEC 10646:2003 に対して追補 1，追補 2 の開発作業が進行中であるが，追加予定となっているのは，インド系文字の起源の 1 つであるカローシュティ文字，アルファベットの直接の祖先であるフェニキア文字，シュメール・アッカド時代の楔形文字，チンギスハン時代のモンゴル帝国が創作したパスパ文字といった歴史上の文字や，インドネシアのスラウェシ島で今も使われているブギス文字，北アフリカのベルベル語話者が使用しているティフナグ文字といった少数民族文字である．新規文字に関する符号開発のプロセスを加速するために，カリフォルニア大学バークレーの言語学者 Deborah Anderson は，2002 年に Script Encoding Initiative という運動を始めた．これは後に Universal Encoding Initiative と名称を変え，2004 年には UNESCO のスポンサーも得て，16 種類の新規文字コード開発に取り組むと宣言している．

筆者は印刷史研究家の小宮山博氏が所蔵する約 130 年前のウィーン王立印刷所の活字見本帳[8] を拝見したことがあるが，そこには，アショカ王碑文体のブラフミ文字，エチオピア文字，アホム文字，アルバニア文字，グプタ朝文字から始まってタガログ文字，タミル文字，テルグ文字，チベット文字，パスパ文字に至る 74 種類の活字見本が収録されていた．これらの中には，依然として UCS に収録されていない文字すらある．今日の情報技術が，多様な文字文化の表現において 130 年前の活字印刷の水準にすら追いついていないというのは残念なことである．すでに忘れ去られた文字文化（**図 -4**）であっても，その再現のために情報技術が活用されることは，文字通り情報技術の恩恵と呼ぶにふさわしい．

現在では，両側のデザインが民族の文字，タガログ文字をモチーフにしたものであることに気づくフィリピン人は少ない．

図 -4　フィリピンで発行されたタガログ文字をデザインした切手

○ おわりに ○

新しい世代を迎えた情報技術が旧世代技術の慣習を継承する現象はしばしば見られる．文字符号もまた旧世代技術のさまざまな慣習，遺産を継承している．ASCII における 7F のコード位置が DEL を意味するのは紙テープの遺産であるし，また，ASCII の 3 列目と 4 列目において，《1》と《！》，《4》と《＄》などが並んでいるのは，英文タイプライタにおける上段と下段のペアを継承したものである．本稿においても，インド系文字におけるタイプライタ時代の慣習がローカルベンダによる多数の文字符号へと継承されている姿を指摘した．こうした慣性の力と後方互換への強い要請があることから，文字符号の決定は多くの場合歴史上 1 回きりの決定であり，いったん決定された符号はディジタル空間における「見えない文字の正書法」として未来を拘束する．しかし，文字符号の決定は当該言語のネットワーク上での利用にあたっての第 1 段階に過ぎない．当該言語・文字の利用が全面的に開花するためにはさまざまな課題が解決されなくてはならない．

筆者らは，マレーシア，インド，タイなどのパートナーとともに，本年 11 月に開催される世界情報社会サミット（WSIS, World Summit on the Information Society）に提出される UNESCO 統計研究所の報告書 "Language Diversity on the Internet: various points of view on the subject" に，"Language Diversity on the Internet: An Asian View" と題して寄稿した．その主旨はネットワーク上でのバランスのとれた言語活動の成長をフォローするために言語天文台の活動が必要である，と主張したものである．筆者らは，文字符号の利用という視点から，言語天文台の活動を通じて本稿で述べた主題の推移を観察し，報告していきたい．

参考文献・参考 URL
1）伊藤英俊：日本語情報処理の諸相：文豪，JIPS, M 式入力などの日本語情報処理開発，情報処理，Vol.45, No.1, pp.68-75 (Jan. 2004).
2）Priolkar, A. K.: The Printing Press in India - Its Beginning and Early Development, Marathi Samshodhana Mandala, Bombay, pp.13-14 (1958).
3）ISO-IR, http://www.itscj.ipsj.or.jp/ISO-IR/
4）IANA Registry of character codes, http://www.iana.org/assignments/character-sets
5）Mikami, Y., Zavarsky, P. et al.: The Language Observatory Project (LOP), WWW2005, Chiba (May 10-14).
6）周　冰洋，刘　檀婷，姚　世全（編）：常用漢字編碼字典，宇航出版社 (1990).
7）芝野耕司：JIS X 0221 (ISO/IEC 10646) の目指すもの―文字コードと日本の国際対応，情報処理学会情報規格調査会 NEWS LETTER, Vol.40 (1998).
8）Alfabete des Gesammten Erdkreises aus der K. K. Hof- und Staatsdruckerei in Wien, Zweite Auflage, Wien (1876).

（平成 17 年 8 月 7 日受付）