失われた文字コード

安岡　孝一（やすおか　こういち）

■ はじめに

「失われた文字コード」とタイトルに付けたものの、実は、文字コードそのものが失われているわけではない。失われているのは、ある文字コードを読むための環境であって、その文字コードで書かれたファイルそれ自体は、古い CD-R や磁気ディスクやその他の記憶媒体の中に、ひっそりとしかし厳然と存在していたりする。そのような「失われた文字コード」で書かれたファイルが、何かのはずみで我々の前に姿を現す。実際、読めない文字コードで書かれたファイルほど、始末の悪いものは無い。しかも、完全に読めないならまだしも、平仮名や片仮名など一部が読めるのに他が文字化けしていたりすると、何としても残りも読みたくなってしまうのが人のサガである。

本稿では、そのような「失われた文字コード」のうち、本誌の読者の手元に残っていそうな文字コードとして、特に「シフト GB」と「文字鏡コード」を取り上げる。また、これらに加えて「住基コード」に対しても、警鐘を鳴らしておくことにする。

■ シフト GB

GB 2312[1] に収録された簡体字を、日本語版 MS-DOS や日本語版 Windows 3.x で使うために開発された文字コードである。本来 GB 2312 は、94 × 94 の文字表を A1A1 ～ FEFE に割り当てて用いるが、日本語版 MS-DOS では FD や FE のコードは使えなかった。そこで、JIS X 0208 をシフト JIS に変換するやり方をまねて、GB 2312 の 94 × 94 の文字表を 47 × 188 に折り曲げ、1 バイト目に 81 ～ 9F・E0 ～ EF を、2 バイト目に 40 ～ 7E・80 ～ FC を使うことで、いわばシフト JIS の上に GB 2312 を載っけてしまうのが、シフト GB の基本的アイデアである。

当然、同じ文字コード上にシフト GB とシフト JIS が重なることになり、同じ「889F」という文字コード（図 1・2 の 16-01 にあたる）であっても、シフト GB では「啊」を、シフト JIS では「亜」を、それぞれ意味することになる。たとえば

図 1　GB 2312 の 16 区（シフト GB の 889F ～ 88FC）

16区	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
01-19		啊	阿	埃	挨	哎	唉	哀	皑	癌	蔼	矮	艾	碍	爱	隘	鞍	氨	安
20-39	俺	按	暗	岸	胺	案	肮	昂	盎	凹	敖	熬	翱	袄	傲	奥	懊	澳	芭
40-59	捌	扒	叭	吧	笆	八	疤	巴	拔	跋	靶	把	耙	坝	霸	罢	爸	白	柏
60-79	百	摆	佰	败	拜	稗	斑	班	搬	扳	般	颁	板	版	扮	拌	伴	瓣	半
80-94	办	绊	邦	帮	梆	榜	膀	绑	棒	磅	蚌	镑	傍	谤	苞	胞	包	褒	剥

図 2　JIS X 0208 の 16 区（シフト JIS の 889F ～ 88FC）

16区	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
01-19		亜	唖	娃	阿	哀	愛	挨	姶	逢	葵	茜	穐	悪	握	渥	旭	葦	芦
20-39	鯵	梓	圧	斡	扱	宛	姐	虻	飴	絢	綾	鮎	或	粟	袷	安	庵	按	暗
40-59	案	闇	鞍	杏	以	伊	位	依	偉	囲	夷	委	威	尉	惟	意	慰	易	椅
60-79	為	畏	異	移	維	緯	胃	萎	衣	謂	違	遺	医	井	亥	域	育	郁	磯
80-94	一	壱	溢	逸	稲	茨	芋	鰯	允	印	咽	員	因	姻	引	飲	淫	胤	蔭

「白雪皑皑」という文字列は、シフト GB では「88D5 9948 88A6 88A6」という文字コードになるが、これをシフト JIS だと思って表示すると「易僣姶姶」となってしまうわけだ。逆に言えば、「易僣姶姶」とシフト JIS で書かれたファイルを、シフト GB のフォントで表示すれば、見事「白雪皑皑」になるという仕掛けだった。これにより、フォント切り換えの可能なワープロソフト上では、日本語と中国語を混在して表示できたのである。

✽ Windows 98 以降のシフト GB

ところが Microsoft は、Windows 98 において全てのフォントを Unicode 化することを決定、同時に MS-Word の内部コードを Unicode 化したため、それまでのシフト GB は使えなくなってしまった。シフト JIS の文字コードのうち、CP932[2] において Unicode との対応を規定されている部分しか、使えなくなってしまったのである。たとえば、CP932 は 9873 ～ 989E（図 3 の 47-52 ～ 47-94）に、文字を収録していない。一方シフト GB は、この部分に「显」～「晓」の 43 字（図 4 の 47-52 ～ 47-94）を収録していた。この結果、Windows 98

47区	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
01–19			蓮	連	錬	呂	魯	櫓	炉	賂	路	露	労	婁	廊	弄	朗	楼	榔
20–39	浪	漏	牢	狼	篭	老	聾	蝋	郎	六	麓	禄	肋	録	論	倭	和	話	歪
40–59	賄	脇	惑	枠	鷲	亙	亘	鰐	詫	藁	蕨	椀	湾	碗	腕				
60–79																			
80–94																			

図 3　JIS X 0208 の 47 区（シフト JIS の 9840 ～ 987E・9880 ～ 989E）

47区	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
01–19			稀	息	希	悉	膝	夕	惜	熄	烯	溪	汐	犀	檄	袭	席	习	媳
20–39	喜	铣	洗	系	隙	戏	细	瞎	虾	匣	霞	辖	暇	峡	侠	狭	下	厦	夏
40–59	吓	掀	锨	先	仙	鲜	纤	咸	贤	衔	舷	闲	涎	弦	嫌	显	险	现	献
60–79	县	腺	馅	羡	宪	陷	限	线	相	厢	镶	香	箱	襄	湘	乡	翔	祥	详
80–94	想	响	享	项	巷	橡	像	向	象	萧	硝	霄	削	哮	嚣	销	消	宵	淆

図 4　GB 2312 の 47 区（シフト GB の 9840 ～ 987E・9880 ～ 989E）

13区	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
01–19															显	险	现	献	县
20–39	腺	馅	羡	宪	陷										限	线	相	厢	镶
40–59	香	箱	襄	湘	乡	翔	祥	详	想	响	享								
60–79					项	巷	橡	像	向	象	萧	硝	霄	削	哮		嚣	销	消
80–94	宵	淆			晓	黠				黟	鼷								

図 5　Unicode 対応シフト GB の 13 区（8740 ～ 877E・8780 ～ 879E）

表 1　シフト GB の Unicode 対応

オリジナルのシフト GB		Unicode 対応シフト GB		
GB 2312	シフト GB	移動先	NEC 拡張文字版シフト GB	IBM 拡張文字版シフト GB
47-52 ～ 47-61	9873 ～ 987C	13-11 ～ 13-20	874A ～ 8753	874A ～ 8753
47-62 ～ 47-77	987D ～ 987E・9880 ～ 988D	13-32 ～ 13-47	875F ～ 876E	875F ～ 876E
47-78 ～ 47-80	988E ～ 9890	13-63 ～ 13-65	877E・8780 ～ 8781	877E・8780 ～ 8781
47-81	9891	13-66	8782	FA59
47-82	9892	13-67	8783	8783
47-83	9893	13-68	8784	FA5A
47-84 ～ 47-88	9894 ～ 9898	13-69 ～ 13-73	8785 ～ 8789	8785 ～ 8789
47-89 ～ 47-93	9899 ～ 989D	13-75 ～ 13-79	878B ～ 878F	878B ～ 878F
47-94	989E	13-83	8793	8793
84-01 ～ 84-94	EA9F ～ EAFC	89-01 ～ 89-94	ED40 ～ ED7E・ED80 ～ ED9E	FA5C ～ FA7E・FA80 ～ FABA
85-01 ～ 85-94	EB40 ～ EB7E・EB80 ～ EB9E	90-01 ～ 90-94	ED9F ～ EDFC	FABB ～ FAFC・FB40 ～ FB5B
86-01 ～ 86-94	EB9F ～ EBFC	91-01 ～ 91-94	EE40 ～ EE7E・EE80 ～ EE9E	FB5C ～ FB7E・FB80 ～ FBBA
87-01 ～ 87-78	EC40 ～ EC7E・EC80 ～ EC8E	92-01 ～ 92-78	EE9F ～ EEEC	FBBB ～ FBFC・FC40 ～ FC4B
87-79	EC8F	13-84	8794	8794
87-80	EC90	13-88	8798	8798
87-81 ～ 87-90	EC91 ～ EC9A	92-81 ～ 92-90	EEEF ～ EEF8	FA40 ～ FA49
87-91	EC9B	13-89	8799	8799
87-92 ～ 87-94	EC9C ～ EC9E	92-92 ～ 92-94	EEFA ～ EEFC	FA55 ～ FA57

やWindows 2000では、シフトGBの「显」～「晓」を表示できなくなってしまった。

この問題を回避するため、シフトGBを使用していた各社は、これらの文字を13区に移動させることにした（図5）。CP932の13区にあたる部分（8740～877E・8780～879E）は、NEC特殊文字と呼ばれる独自拡張文字を収録していたが、この部分をこれまでのシフトGBは使っていなかったからだ。また、GB 2312の84区以降に関しても、NEC拡張文字のある89区以降に移動させた。移動の詳細を表1に示す。なお、NEC拡張文字（ED40～EEFC）に対しては、全く同じ文字が多少順序を変えてIBM拡張文字（FA40～FC4B）にも収録されており、ファイル入出力の際に相互に自動変換される。このような自動変換がおこなわれた場合の文字コードを、表1に「IBM拡張文字版シフトGB」として示しておく。

✽ シフトGBのサルベージ

前世紀末には一世を風靡したシフトGBも、今世紀に入ってからは衰退の一途をたどっている。開発の一翼を担ったオムロンも、2010年5月で『楽々中国語』（cWnn）のサポートを完全に打ち切ってしまった。現時点でシフトGBをサポートしている製品は、わずかに高電社の『ChineseWriter』のみ[3]となってしまっている。では、何かの拍子にシフトGBのファイルを発掘してしまった場合、『ChineseWriter』を購入して読むしかないのだろうか。

実は、シフトGBのサルベージに関しては、針谷壮一氏が公開している「中国語コンバータ[4]」というフリーウェアが、現時点において最も強力なツールである。各社固有のシフトGBテキストをGB 2312のA1A1～FEFEに変換してくれると同時に、シフトGB・シフトJIS混在のrtf形式も変換可能というスグレモノだ。すなわち、doc形式のシフトGBが発掘された場合でも、MS-Wordの「リッチテキスト形式」でそのまま保存しなおすことができれば、「中国語コンバータ」によって、フォント名も含め変換可能[5]となっている。

ただし、サルベージ変換後のテキストやrtfは、現在は読み書きが出来ても、いずれ遠い将来には、また「失われた文字コード」となりかねない。これらのテキストやrtfは、プリントアウトして紙の形にしておくか、せめてフォント埋め込みのPDFにしておく必要があるだろう。

■ 文字鏡コード

『今昔文字鏡』で使われている10進6桁の文字鏡番号に対し、000001～005640、005641～011280、011281～016920、……、と5640個ずつに区切り、各5640個をJIS X 0208の16～45・48～77区に割り当てた文字コードである。文字コード上は完全にダブった形となるが、000001～005640は「Mojikyo M101」フォントに、005641～011280は「Mojikyo M102」フォントに、011281～016920は「Mojikyo M103」フォントに、……、それぞれ収録することにより、フォント切り換えによって全ての文字を扱うことができる、というのがウリであった。逆に言えば、文字鏡フォントが実装されていない環境下では、文字鏡コードで書かれた「一」「声」「鴆」……は、全て「亜」に文字化けすることになる（図6～8および図2）。

文字鏡番号の000001～049964は基本的に『大漢和辞典』の検字番号を踏襲しているが、それ以降の番号は文字鏡研究会が独自に付けたものである。管理を担当する文字鏡フォントセンターは、2000年9月の使用許諾書において、文字鏡番号と他の文字集合との対応表および、その対応表を用いて文字の変換をおこなうソフトウェア等の作成・配布を、明確に禁止している。

✽ 文字鏡コードの現況

文字鏡コードで書かれたMS-WordやExcelあるいは一太郎などのファイルは、文字鏡フォントが実装されていない環境では文字化けが起こるが、文字鏡フォントセンターは2008年3月以降、文字鏡フォントの無償ダウンロードを打ち切っている。文字鏡番号の付与をおこなってい

たはずの文字鏡研究会は、ほぼ同時期に活動を凍結し、そのWWWサイト[6]も「再構築中」の状態が続いている。また、現時点で文字鏡コードをサポートしている製品は、エーアイ・ネットの『今昔文字鏡』のみである。ただし、各文字鏡番号に対応する字形に関しては、ISO/IEC 10036のグリフ登録サイト[7]で閲覧可能だ。ISO/IEC 10036のグリフ番号10000001～10576000が文字鏡番号000001～576000に対応しており、128ドットのgif画像が入手できる。

このような状況において、文字鏡コードで書かれたファイルが発掘された場合、どうやってサルベージすればいいのか。残念ながら文字鏡フォントセンターが、そのようなサルベージをおこなうソフトウェアを明確に禁止してきたため、現時点では全く妙案がない。化けてしまった一字一字に対して、その字のJIS X 0208における区点番号を調べて、フォント名と共に文字鏡番号に変換[8]し、さらに10000000を加えて、ISO/IEC 10036のグリフ登録サイトからgifを入手[9]するしかない。もし、フォント名がわからない場合は、とりあえず「Mojikyo M101」の文字鏡番号を計算し、そのグリフ番号に順に5640を加えていって当該グリフを入手、というのを繰り返すことになる。こんなのを手作業でおこなうのは、全く馬鹿げているとしか言いようがないのだが、ソフトウェアを作成できない以上、どうしようもない。

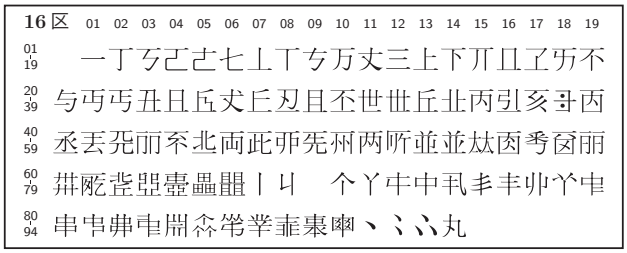

図6 「Mojikyo M101」の16区（文字鏡番号000001～000094）

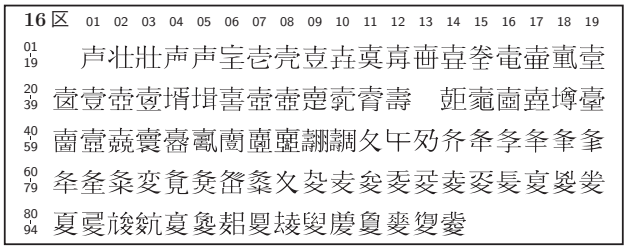

図7 「Mojikyo M102」の16区（文字鏡番号005641～005734）

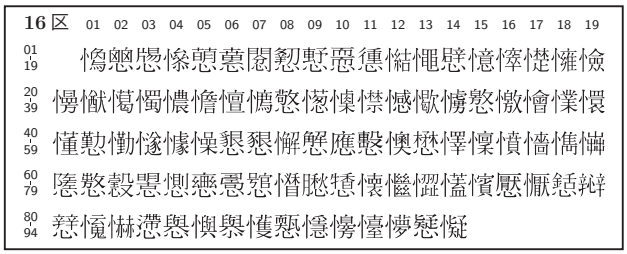

図8 「Mojikyo M103」の16区（文字鏡番号011281～011374）

■ 住基コード

住民基本台帳ネットワークにおける統一文字コードとして、2001年2月に検討版が作成され、2001年10月に制定、2002年8月より運用開始された文字コードである。Unicodeの基本多言語面（U+0000～U+FFFF）を独自拡張した2バイトの文字コードであり、0000～9FFFおよびE000～FFFFに関してはUnicodeと互換である（図9）。ただし、ハングルにあたる領域（AC00～D7A3）に関しては、ハングルではなく独自の拡張漢字や変体仮名などを、勝手に割り当てる形で設計された（図10）。これに加え制定版では、JIS X 0213の第3・4水準漢字のうち漏れていた漢字を、AAA1～ABCFに追加している（図11）。

住基コードの仕様は一般には非公開となっており、現時点で唯一のフォントである「KAJO J明朝」を購入して、文字コードを調べるしかない。ただし、2009年3月発行の『汎用電子情報交換環境整備プログラム成果報告書別冊』（日本規格協会）には、住基コードの漢字部分が収録されており、漢字に関しては一応の調べがつくようになった。

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
4E0x	一	丁	丐	七	丄	丅		万	丈	三	上	下	丌	不	与	丏
4E1x	丐	丑	丒		且	丕	世	丗	丘	丙	业				丞	丟
4E2x		両	丟	丧	两		並		丨	丩	个	丫	丬	中	丮	丰
4E3x	丰	丱	串			丳	丶		丸	丹		主	丼			丿

図9　住基コードの4E00～4E3F

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
AD0x	亍	万	五	丈	兀	丈	左	刃	丑	刃	丑	才	廿	丙	充	丼
AD1x	弐	丼	盯	所	亟	亟	垢	乖	砝	爾	丰		暢	々	主	来
AD2x	乂	乂	乂	久	凡	乕	乳	乳	亂	豫	矣	亥	交	亮	亭	亮
AD3x	牽	亶	齊	襄	齊	廖	舗	齋	齋	齋	廖	亹	齎	亹	變	个

図10 住基コードのAD00～AD3F

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
AAAx		亠	丈	斥	自	入	倆	俚	僕	儡	儷	儼	兇	浴	八	剱
AABx	劉	劭	勵	斗	卓	云	艮	咤	晷	嘘	嚠	嘑	噬	圍	土	坑
AACx	垢	坤	埆	垈	堅	垓	埈	堀	塑	壅	壇	夷	奠	鉄	姶	姚
AADx	嬬	宇	屎	屹	妄	岿	岷	岌	峇	岸	岼	嵜	峻	嶢	嶄	

図11 住基コードのAAA1～AADF

✽ 住基コードの崩壊

2008年12月にTai Viet[10]がU+AA80～U+AADFに収録されたことから、住基コードは急速に崩壊が進んでいる。というのも、たとえばU+AABEは「TAI VIET VOWEL AM」という非前進文字なので、通常のUnicode処理では直前の文字に重ねて表示される。この結果、住基コードのAABEにあたる「丄」は、直前の文字と重なって表示されてしまう[11]のである。

今後、UnicodeのU+AAA1～U+ABCFに新たな文字が追加されるたび、住基コードの崩壊が加速していくのは間違いない。本来、これらの独自拡張漢字は、U+20000～U+2FFFFの拡張漢字面にちゃんと収録するか、さもなくばIVS[12]を使って見分けるべきなのだ。元々の設計が完全に誤っている上に、住基コードの仕様を公開しないという方針を堅持し続けているため、もはや誰からも救いの手を差しのべようがない、というのが現状である。

■ おわりに

シフトGB、文字鏡コード、住基コード、の3つの文字コードについて概要を述べ、これらの文字コードがどのようにして「失われていった」のかを解説した。これらの文字コードに共通する特徴は、特定のベンダやグループによってややもすれば閉鎖的に使われており、標準化からは程遠い立場にあったことである。言い換えれば、標準化をおこたった文字コードは、いずれ「失われる」ということだろう。

もし、読者諸氏が今現在、仕様が一般に公開されていない文字コードを使っているのなら、悪いことは言わない、さっさと別の文字コードに乗り換えるべきだ。閉じた仕様の文字コードは、いずれ誰にもサポートできなくなるのだから。その点を最後に指摘して、本稿をしめくくることにする。

注

[1] 中華人民共和国の国家標準局が1980年に発布した漢字コード。94×94の文字表の中に、非漢字682字、第1級漢字3755字、第2級漢字3008字、あわせて7445字を収録していた。現在はGB 18030に、とって代わられている。

[2] Microsoftの独自拡張シフトJIS。

[3] 高電社版のシフトGBである「CWコード」では、1～8区の非漢字部分はGB 2312ではなくJIS X 0208である。すなわち、非漢字部分はシフトJISで、漢字部分だけがUnicode対応シフトGBだと言える。

[4] http://www5b.biglobe.ne.jp/~harigaya/chcnv.html

[5] 『ChineseWriter』の「CW-GB Mincho」や「GB中国

明朝」、『楽々中国語』の「OM 中国 GB 明朝」、あるいは『NiHao2』の「NH 簡体宋」などのフォントを、Microsoft Windows の「SimSun」に変換し、同時に文字コード変換をおこなう。ゴシックなども同様。

[6] http://www.mojikyo.org/

[7] http://10036ra.org/glyph-index.html

[8] 16 ～ 45 区の場合は、区番号から 16 を引いて 94 倍し、点番号を加える。48 ～ 77 区の場合は、区番号から 18 を引いて 94 倍し、点番号を加える。さらに、フォント名の番号部分（「Mojikyo Mxxx」の xxx の部分）から 101 を引いて 5640 倍した値を加えれば、文字鏡番号が得られる。

[9] グリフ番号 10050001 の gif が http://10036ra.org/retrieve.php?gid=10050001 で入手可能。

[10] 黒タイ語や白タイ語に用いられる文字。

[11] U+2123D に収録されている本物の「土」を使えば、このような重なり表示の問題は発生しないが、住基コードでは U+10000 以降を使うことができない。

[12] Ideographic Variation Sequence。Unicode において複数の字体を見分けるために付加する文字コード。たとえば「丈」と「丈」は、それぞれ「U+4E08 U+E0100」と「U+4E08 U+E0101」という IVS で見分けられる。