

この PDF ファイルは、Dirk Meyer 氏による「Summary, Explanations, And Remarks: GB 18030-2000」([ftp://ftp.oreilly.com/pub/examplesnutshell/cjkv/pdf/GB18030\\_Summary.pdf](ftp://ftp.oreilly.com/pub/examplesnutshell/cjkv/pdf/GB18030_Summary.pdf)) の 1.4 版を直井 (y.naoi@glamour.co.jp) が日本語訳したものであり、原著者の許可を得た上で公開するものです。

テキストの日本語化にあたっては、もろしげき氏、野村英登氏、小形克宏氏をはじめ漢字文献情報処理研究会 (<http://www.jaet.gr.jp/>) 有志のご協力をいただきましたが、誤訳があれば 100 パーセント直井のせいです。日本語版テキストについてご指摘や疑問点などがありましたら、Meyer 氏ではなく、直井までご連絡いただきますようお願いいたします。

以下、左の欄に記されているのはすべて訳註です。

San José, February 4, 2001

本稿 (version 1.4) は、筆者のメールアドレス (dmeyer@adobe.com) を含めて再配布自由とする。また、訂正すべき点や追加情報があれば、筆者に連絡していただきたい。本稿は GB18030-2000 の直訳ではなく、この規格を理解するための一助であると考えていただきたい。

テキストの随所に挿入されている [ ] は、中国で発行された規格票の対応するセクションを示ものである。

本稿に対して最初の批評をしてくれた Markus Scherer 氏 (IBM) と Ken Lunde 氏 (Adobe Systems) に感謝申し上げます。

#### 主な変更点とマッピング・データ情報

本稿を最初に執筆した後に、GB 18030「符号化文字集合」と Unicode とのマッピングについて、相当数の変更が発表された。この変更の最大の帰結は、U+0080 からはじまる Unicode BMP 符号位置に対応するすべての 4 バイト GB シーケンスが、完全に並べ直されたことであろう。

マッピング・データは 2000 年の末に再発行されたが、これはマッピング・データおよびそれに関連する大きな区分に限って変更を反映したものである。このマッピング・データに基づいて、本稿に手短な更新情報を追加した。該当箇所は、**now:** や **UPDATE:** という注記で変更点を示している。ただし、規格の履歴を継続的に記述するために、しばしば古い値と訂正された値が隣り合って混在することとなる。正確な更新済みの情報を得るには、最新のマッピング・データ (以下に示す) を参照していただきたい。

現在開発者は、このような変更について明確な記載があるはずの GB 18030 第 2 版を待ち望んでいる。この第 2 版では、規格自体の構造上の原則については有効なままであると予想されるが、一方で、マッピング・データがらみのその他の変更点が明らかになるであろう。本稿の多くの部分は、いまなお規格第 1 版を参照している。これらの部分は、新版が入手可能となるまで変更されることはないであろう。いくつかの部分が十分な明確さを欠いているとすれば、おそらく規格第 2 版で提供されるはずの必要な説明と定義を持たないことによるものであろう。

現在および将来の、GB 18030 と Unicode とのマッピングの経過を追うための優れた資料は、IBM/Cupertino のサーバ上に設置されているオープンソース「International Components for Unicode」計画のウェブ・サイトである。この ICU サイト (<http://oss.software.ibm.com/icu/charset>) では、XML 形式のダウンロード可能なマッピング・データを見出すことができる。[ICU 計画については、<http://oss.software.ibm.com/icu/index.html> を参照されたい]

## GB 18030-2000 の要約と解説, 所見

CHINESE NATIONAL STANDARD GB18030-2000: INFORMATION TECHNOLOGY – CHINESE IDEOGRAMS CODED CHARACTER SET FOR INFORMATION INTERCHANGE – EXTENSION FOR THE BASIC SET  
(信息技术—信息交换用汉字编码字符集—基本集的扩充 *Xinxi jishu – Xinxi Jiaohuan Yong Hanzi Bianma Zifuji – Jibenji De Kuochong*)

中国の国家規格 (国家标准 *guojia biao zhun*) である GB 18030-2000 (以下, GB 18030) が, 2000年3月17日に発行された。この規格は, Unicode version 3.0 の出現によって発生する課題の解決をはかるものである。具体的には, Unicode の拡張された文字レパートリ, すなわち統合漢字 Extension A を, 中国の過去の国家規格が網羅している文字とともに使えるようにしようとするものである。

### 沿革

中華人民共和国はすでに, ISO/IEC と Unicode Consortium の共同成果を支持する基本合意を表明し, 符号および文字において ISO 10646-1/Unicode 2.1 と互換性を持つ国家規格を発行している。この規格が GB 13000.1 であり, ISO や Unicode Consortium がその「共通の」規格を変更または改訂した場合には, GB 13000.1 にもその変更が即座に反映されることとなっていた。

しかし Unicode/GB 13000.1 が発行されたときには, 中国の簡体字を表記するための国家規格である GB 2312 がすでに存在し, 広く使われていた。そこで, GB 2312 との互換性を維持するために, GBK という「仕様書」が創案されたのである。GBK は GB13000.1 とほぼ共通の文字レパートリを持つ第2の符号化文字集合だが, まったく異なる符号化を用いる。GBK は “*Guojia biao zhun kuozhan*” の略称である。正式名称は「*Hanzi neima kuozhan guifan* 汉字内码扩展规范」あるいは “Rules/Specifications defining the extensions of internal codes for Chinese ideograms” である (原註: 本稿では, 中国語の「汉字」を「Chinese ideograms」と訳す)。

GBK の際立った特徴は, GB 2312 で定義されている文字と符号を一切変更することなく, すべての追加文字をその周辺に配したことである。追加文字は主として, Unicode 2.1 の統合漢字部分——これは GB 2312 の文字レパートリよりも大規模である——の文字である。したがって GBK では, GB 2312 との間で文字と符号の互換性が確保され, 同時に完全な Unicode 統合漢字の文字集合が利用可能となる。GBK の制定にあたっては, 当時の Unicode には収録されていなかった文字も追加された。

われわれはまず, GBK の符号空間と文字レパートリを見ていくこととする。これは, ほぼ完全な互換性を有する GB 18030 の符号空間の基盤をなすものである。事実, 規格票には, GB 18030 (「規格」) は GBK (「仕様書」) を置き換える (代替 *daiti*) ことになるという記述がある。

GBKの2バイト符号空間は、以下のように定義される。

記号用標準領域（符号标准区 *fuhao biao zhun qu*):

- GBK/1: 0xA1A1-0xA9FE (846 符号 / 717 記号)
- GBK/5: 0xA840-0xA9A0 (192 / 166 記号)

漢字用標準領域（汉字标准区 *Hanzi biao zhun qu*):

- GBK/2: 0xB0A1-0xF7FE (6,768 / 6,763 漢字)
- GBK/3: 0x8140-0xA0FE (6,080 / 6,080 漢字)
- GBK/4: 0xAA40-0xFEAO (8,160 / 8,160 漢字)

ユーザ定義領域（用户自定义区 *yong hu zi ding yi qu*):

- UDA 1: 0xAAA1-0xAFFE (564 / 0)
- UDA 2: 0xF8A1-0xFEFE (658 / 0)
- UDA 3: 0xA140-0xA7A0 (672 / 0)

このようにGBKは、21,886文字を含む23,940の符号位置を定義している。同時にGBKは、Unicode 2.1の符号位置へのマッピングを提供する。GBK発行時点でUnicodeに含まれていなかった文字は、符号位置U+E000からはじまるUnicodeの私用領域(PUA)との対応を与えられている。GBKの空き符号位置についても同様である。

GBKの定義に使われて満杯になった符号空間を見ると、大規模な追加のために残された空間がないことが明らかになる。GBKの3つのユーザ定義領域の1,894の符号位置では、統合漢字Extension Aのために十分な空間を提供する道さえ閉ざされている。というのも、Extension Aは6,582の新しい文字をUnicode 3.0の第0面、基本多言語面(BMP、この用語はISO 10646に起源を持つが、現在ではUnicodeの文脈でも用いられる)に定義しているのだ。

## GB 18030で拡張された点

GB 18030は定義上「符号化文字集合」(编码字符集, *bian ma zi fu ji*)であり、文字レパートリのみならず文字の符号位置をも規定するものである。

GB 18030には、以下のような特徴がある。

- GB 18030は、Unicodeの統合漢字Extension Aを完全に包含する。
- GB 18030は、GBKにすでに含まれている符号位置に加え、Unicodeの第0面(BMP)および他の16の面の全符号位置に対して——未定義の部分も含めて——符号空間を提供する。換言するならGB 18030はGBKに対して符号および文字の互換性を有する「上位集合」でありながら、同時に、GBKに含まれていないすべてのUnicode符号位置に対して空間を提供しようとするものである。それゆえ事実上、GB 18030の部分集合とUnicodeの全符号化空間のあいだには、1対1の対応が発生する。
- 統合漢字を組み込み、Unicode 3.0に対して符号空間を割り当てるために、GB 18030は4バイト符号化機構を定義し利用している。

上の第2点で述べたような符号の対応付けを示すために、GB 18030でこれがどのように実現されているかの例を挙げる。

- 規格票には Unicode と GB 18030 の対応を示すテーブルが収録されているが、われわれはそこで 0x81308434 (**now:** 0x81308435) と 0x81308435 (**now:** 0x81308436) の間に、Unicode 符号マッピングの第1の欠落を見出すことになる。飛び越されたのは U+00A4 (通貨記号) である。われわれはこの文字を GB 18030 の符号位置 0xA1E8 (この文字は、GB 2312, GBK, そして GB 18030 のいずれにおいても、この位置に置かれている) で発見するであろう。

- さらに2つのより特徴的な欠落を、Unicode に対応付けられる GB 18030 の4バイト領域において確認することができる。予想できることであるが、この欠落のうちの1つは、Unicode のCJK統合漢字レパートリがGBK/GB 18030の主要部分であるという事実を反映しており、U+4DFF は 0x82358F33 (**now:** 0x82358F32) に、U+9FA6 は 0x82358F34 (**now:** 0x82358F33) に対応付けられる。

もう1つの特徴的な対応上の欠落は、Unicode の私用領域 (PUA) に関連したものである。PUA の直前の Unicode 符号位置である U+DFFF は 0x83389837 に対応し (**now:** U+D7FF が 0x83389838 に対応し、サロゲート領域へのマッピングはどこにも存在しない)、GB 18030 の非4バイト領域に対応位置を持たない最初の PUA 符号位置である U+E865 が、0x83389838 (**now:** 0x8336D030) に対応する。この間の PUA 符号位置は、GB 18030 の以下の領域に見出せる。

U+E000-U+E765: 2バイト (ユーザ) 領域 1, 2, 3

(0xAA-0xAF, 0xF8-0xFE, 0xA1-0xA7)

U+E766-U+E7BB: 2バイト領域 1 (0xA1-0xA7)

U+E7BC-U+E7C6: 2バイト領域 5 (0xA8)

U+E7C7-U+E7E1: 2バイト領域 1 (0xA8)

U+E7E2-U+E7FD: 2バイト領域 5 (0xA9)

U+E7FE-U+E80F: 2バイト領域 1 (0xA9)

U+E810-U+E814: 2バイト領域 2 (0xD7)

U+E815-U+E864: 2バイト (ユーザ) 領域 4 (0xFE)

**UPDATE:** 現在、再発行されたマッピング・データには、第3の大きな欠落を見出すことができる。これは、4バイト領域と Unicode のサロゲート符号位置 (U+D800 から U+DFFF) との対応付けが削除されたことによる。

## GB 18030 の主要部分の要約

規格票前文は、以下のように述べている。

- GB 18030 は、GB2312 を拡張するものである。
- GB 18030 は、GBK 1.0 版を置き換える (代替 *daiti*) ものである。
- GB 18030 は、中華人民共和國情報産業省が公布するものである。
- 以下の研究機関および企業が、規格原案作成に参画した。

情報産業省電子工業標準化研究所  
北京大学コンピュータ技術研究所  
北大方正グループ  
北京方正新天地  
四通グループ

中科院ソフトウェア研究所  
長城ソフトウェア  
四通利方  
中軟総、金山ソフトウェア、联想  
言及されている個人参画者は、以下の通り。  
陳堃鈺、黄疆、胡万進、張建国、陳壯 [前文]

規格票本文は、以下を定義している。

1) 引用規格

- GB 2311-1990 (ISO 2022:1986 と同等), 「情報処理—7ビット及び8ビット符号化文字集合—符号拡張法」
- GB 2312-1980 基本集, 「情報処理—情報交換用漢字符号化文字集合—基本集」
- GB 11383-1989 (ISO 4873:1986 に対応する国際一致規格), 「情報処理—情報交換用8ビット符号—構造および実装規則」
- GB 12345-1990, 「情報交換用漢字符号—補助集」
- GB 13000.1-1993 (ISO/IEC 10646.1-1993 に対応する国際一致規格) 「国際符号化文字集合 (UCS) —第1部 体系及び基本多言語面」 [2]

繰り返すが、GBKは「仕様書」であって「規格」ではないため、このリストでは言及されていない。

2) GB 18030の2つの主導的原則は、それがGB 2312との「符号化規格としての互換性」(内码标准兼容 *neima biao zhun jian rong*) を保持することであり、収集された文字については「GB 13000.1のCJK統合漢字集合およびCJK統合漢字 Extension Aの完全なサポート」である。[3]

3) 用語の定義

- 中国語の用語「字汇 *zihui*」は「レパートリ」を意味する。
- 中国語の用語「字符 *zifu*」は「文字」を意味する。
- 中国語の用語「编码字符 *bianma zifu*」は「符号化文字」を意味する。
- 中国語の用語「保留区 *baoliu qu*」は「未定義領域」を意味する。

4) 文字レパートリ

a) 1バイト領域

- 全128文字。GB 11383で0x00から0x7Fに定義されている文字、および0x80に追加された1バイト版の通貨記号「ユーロ」。(now: 「ユーロ」は、上記の位置から削除され、0x0080は4バイト領域の最初の使用可能な符号位置に割り当てられた)

b) 2バイト領域

- GB 13000.1で定義されている全CJK統合漢字。
- GB 13000.1のCJK互換漢字領域から選定された21の漢字。
- 台湾で用いられる139の記号類。これらは、GB 13000.1に含まれ、GB 2312には含まれない。
- GB 13000.1に含まれるその他31文字<sup>(4)</sup>。
- GB 2312の非漢字。
- GB 12345の19の縦書き用句読記号。
- GB 2312に含まれない10の小文字ローマ数字。

(1) この記述は規格票自体に見られるものだが、「31文字」の内訳は不明。「13文字」の誤り(2バイト領域5の未定義文字まで「文字」として数えてしまうと、31文字になる)か?

(2) この記述は規格票自体に見られるものだが、「GB 2312に含まれないダイアクリティカル・マーク付きピンイン字母」は「4つ」ではないのか？

(3) ここでは原文通り「a」「g」としておいたが、これらの文字に通常のラテン文字とは別の符号位置が与えられているのは、「a」ではなく「a」, 「g」ではなく「g」という意図によるものだろう。

(4) この文字はGB 2312-1980自体には、確かに含まれない。ただし、GB 6345.1 (GB 2312の拡張)で追加されたものであって、一般にGB 2312の実装であると理解されているものの多くがこれらを含む。

(5) 規格票自体にも「GB 11383未採用的 0x30 至 0x39」とあるが、GB 11383の 0x30 から 0x39 には、わたしの理解の範囲では、「0」から「9」の文字が符号化されており、空き領域ではない。たぶんわたしが何か不吉な誤解をしているのだろう。

- 5つのダイアクリティカル・マーク付きピンイン字母<sup>(2)</sup>、および文字「a」「g」<sup>(3)</sup>。これらはGB 2312に含まれない<sup>(4)</sup>。
- 漢数字「ゼロ」。
- 13の漢字記述記号（表意文字描述符 *biaoyi wenzi miaoshu fu*）。
- 追加漢字と部首（部首 *bushou*）または構成要素（构件 *goujian*）。合計80文字。
- 2バイト版の通貨記号ユーロ。

上記のb)で定義した2バイト文字レパートリに関して言えば、GB 18030は実質的にGBKの完全な上位集合である。

#### c) 4バイト領域

- GB 13000.1に存在するCJK統合漢字Extension Aの全漢字。ただし、上述の2バイト領域にすでに符号化されている文字を除く。[5]

この事実は、基本的に符号位置の重複が存在せず、それぞれの文字のための唯一の「登録」位置が確保され、他方、1対nまたはn対1対応が回避されていることを暗示する。

#### 5) 全体的な符号構造

GB 18030は、1, 2, 4バイト文字符号化方式を採用する。

1バイト部分は、GB 11383の符号化構造と原則を適用し、0x00から0x80 (**now:** 0x79)の符号位置を用いる。

2バイト部分は、2つの8ビット・バイナリ列を用いて文字を表現する。第1(先行)バイトは0x81から0xFE, 第2(後続)バイトは0x40から0x7E, および0x80から0xFEの範囲の符号位置となる。

4バイト部分は、GB 11383の空き領域である0x30から0x39の符号位置を用いて2バイト符号を拡張する<sup>(5)</sup>。したがって実際に使用可能な4バイト符号は、0x81308130から0xFE39FE39までの範囲の符号位置となる(表1および図1を参照)。

規格票の表1と図1は、可能な4バイト組成の並びと序列を示しており、それぞれのバイトは、4つの範囲——そのバイトの許可された値を示している——のうちのいずれかとなる。すなわち、第1および第3バイトは0x81から0xFE, 第2および第4バイトは0x30から0x39である。意図された符号シーケンスを得るためには、第4バイト、第3バイト、第2バイト、第1バイトの順で値を増加させていけばよい。[6]

規格票によって定義され、4バイトで表現される符号空間の範囲は以下のとおりである。

```
0x81308130 - 0x81308139,  
0x81308230 - 0x81308239,  
...  
0x8130FE30 - 0x8130FE39,  
0x81318130 - 0x81318139,  
...  
0x8131FE30 - 0x8131FE39,  
...
```

0x82308130 - 0x82308139,  
...  
0x8230FE30 - 0x8230FE39,  
...  
0xFE308130 - 0xFE30FE39,  
...  
0xFE39FE30 - 0xFE39FE39.

以上の原則を適用することで得られる符号位置の数は、GB 18030の1バイト部分で128, 2バイト部分で23,940, そして4バイト部分で1,587,600である。[6]

6) 文字 / 符号位置割り当ての並び順

- a) 1バイト部分の文字とそのシーケンスは、GB 11383のそれぞれの文字と等しい。1バイト通貨記号ユーロは、GB 13000.1の0x20ACを表現するために0x80の位置に追加される(図2を参照)。<sup>[7.1, 8.1]</sup>  
(**now:** 通貨記号ユーロは、この位置から削除された。現在では全角ユーロ 0xA2E3がU+20ACに対応する)
- b) 2バイト部分の文字は、附属書Aに示されているように、順次並べられる。前述のように、この並び順は2バイト通貨記号ユーロを除けば、GBKのそれと実質的に同一であり、2バイト領域1(0xA1A1-0xA9FE)の文字数は、1文字(ユーロ)だけ増加して718文字となる。図3および表2を参照。<sup>[7.2]</sup>

念のため、GB 18030の2バイト符号空間を以下に列挙しておく。

記号用標準領域(符号标准区 *fuhao biao zhun qu*):

- 2バイト領域1: 0xA1A1-0xA9FE (846符号 / 718記号, 0xA2E3に新たにユーロを含み、これはU+E76Cにマップされる。**now:** このユーロは現在はU+20ACに対応する)
- 2バイト領域5: 0xA840-0xA9A0 (192 / 166記号)

漢字用標準領域(汉字标准区 *Hanzi biao zhun qu*):

- 2バイト領域2: 0xB0A1-0xF7FE (6,768 / 6,763漢字)
- 2バイト領域3: 0x8140-0xA0FE (6,080 / 6,080漢字)
- 2バイト領域4: 0xAA40-0xFEA0 (8,160 / 8,160漢字)

ユーザ定義領域(用户自定义区 *yonghu zidingyi qu*):

- ユーザ定義領域1: 0xAAA1-0xAFFE (564 / 0)
- ユーザ定義領域2: 0xF8A1-0xFEFE (658 / 0)
- ユーザ定義領域3: 0xA140-0xA7A0 (672 / 0)

これらの3つの領域は、2つの符号空間——0x8140から0xFE7E, そして0x8180から0xFEFEの範囲——に合計23,940の符号位置を有し、ここに21,887文字が割り当てられている。

より詳細には、GB 18030は、以下のような特殊な文字のグループとその位置について列挙している。

- 2バイト領域2, 3, 4は, 最初にCJK統合漢字を, 続けて追加漢字を格納する。
- GB 2312 で符号化されている漢字は, 2バイト領域2を占める。
- 2バイト領域4の符号位置 0xFD9C から 0xFDA0, および 0xFE40 から 0xFE4Fは, GB 13000.1 から選ばれた21文字のCJK互換漢字を格納する。
- 80の追加漢字と部首・構成要素も, 2バイト領域4に符号化される。
- GB 13000.1 に含まれるがGB 2312に含まれない——台湾で用いられる——139の記号類, 漢数字「ゼロ」, そして13の漢字記述文字は, 2バイト領域5に符号化される。
- GB 2312の非漢字記号, GB 2312に含まれない5つのダイアクリティカル・マーク付きピンイン字母および文字「a」「g」, GB 2312に含まれない10の小文字ローマ数字, GB 12345の19の縦書き用句読記号, 通貨記号ユーロは, すべて2バイト領域1に符号化される。[8.2]

4バイト符号化における文字の並び順は, 以下のとおりである。

- 0x81308130 から 0x8439FE39 の50,400の符号位置は, 2バイト領域に格納されていないGB 13000.1の2バイト文字を符号化し, その並び順はGB 13000.1に従う。残りの符号位置は保留される(剰余碼位保留 *shengyu mawei baoliu*)。
- 0x85308130 から 0x8539FE39 の範囲の12,600の符号位置は, 将来的な文字の拡張用に予約された領域を表現する。
- 0x86308130 から 0x8F39FE39 の範囲の126,000の符号位置は, 漢字の拡張用に予約された領域を表現する。
- 0x90308130 から 0xE339FE39 の範囲の1,058,400の符号位置は, GB 13000.1の16の追加の面を格納するために用いられるであろう。これらの符号位置の並び順は, GB 13000の追加面の符号位置の並び順を完全に忠実に反映したものとなるだろう。残りの符号位置は予約されるであろう。
- 0xE4308130 から 0xFC39FE39 の範囲の315,000の符号位置は, 規格の将来の拡張のために予約された領域を表現する。
- 0xFD308130 から 0xFE39FE39 の範囲の25,200の符号位置は, ユーザ定義領域である。[7.3, 8.3]

#### 7) 附属書

- 附属書Aは, 2バイト文字とそのUnicode値を表にしたものである。
- 附属書Bは, 以前はPUAの符号値を用いて符号化されていた漢字記述文字を掲げている。これらの符号位置は現在, GB 18030の0xA989から0xA995の範囲の2バイト符号領域にある。
- 附属書Cは, 2バイト符号領域の0xFE50から0xFEA0に収録された追加漢字と部首・構成要素を掲げ, そのUnicode値を列挙している。
- 附属書Dは, 4バイト文字領域の完全な範囲, およびこれらの文字に対応するUnicode値を示している。このうち統合漢字 Extension Aによって占められる領域は, 0x8139EF30から0x82358739 (**now:** U+3400/0x8139EE39からU+4DFF/0x82358F32)の範囲の符号空間の文字に対応する(BMPの符号位置の総数は, ここでは6,530であり, これはExtension Aの6,582文字から2バイト文字領域にすでに符号化されている52文字を引いた数と等しい)。

- 附属書 E は、かつて GBK で符号化されていたものの、Unicode の私用領域の符号位置に対応付けられていた文字を列挙している。これらの文字は現行の Unicode 3.0 に含まれるため、その符号化は変更された。

附属書 E 冒頭の解説は、GB 18030 の附属書 A (2 バイト文字のリスト) は——これまで使われてきた GBK との互換性を理由として——いまだに GB 13000.1 のいわゆる「臨時符号」(これらの文字は、GBK に含まれていた) との対応を列挙していると述べている。

(6) 0xA8AC の「小文字 m アクセント」(U+1E3F) は？

影響を受ける文字は、13 の漢字記述文字、「小文字 n グレーブ」<sup>6)</sup>、52 の追加漢字、そして 13 の部首 (now: 14 の部首。Unicode の CJK 部首補遺は U+2E97 を含んでおり、これは GB 18030 の 0xEE5E と等しいと思われる) である。挙げられているのは、それらの GB 13000.1/Unicode における以前の符号位置 (「臨時符号」临时代码 *linshi daima*) および現在の GB 18030/Unicode 3.0 における符号位置である。

## 8) その他所見

### a) 誤植

1) 294 頁 (附属書 E): 0xA987 が「漢字異体字指示子」の符号位置として挙げられているが、正しい位置は 0xA989 のはずである。この点に関して、附属書 B は正しい。

2) 82 頁 (附属書 A): 符号位置 0xA98A に対して、Unicode の U+E7E6 への対応が示されている。これは 0xA95F のための対応付けが重複していることを意味する。0xA98A についての正しい値は、U+E7E8 のはずである (附属書 E を参照)。

3) 11 頁 (附属書 A): 符号位置シーケンス 0xA5FB から 0xA5FD に対して、Unicode の U+E681 から U+E683 への対応が示されている。これは、0xA57B から A57D のための対応付けが重複していることを意味する。0xA5FB 以降についての正しい値は、U+E781 から U+E783 のはずである (附属書 E を参照)。

### b) 異体字指示子の「追加」, 「欠番」の 0x80

GB 18030 の 4 バイト領域から Unicode への対応付けを列挙する附属書 D の 149 頁で、われわれは 0x8139A634 の U+303E へのマッピングを発見した。これは、附属書 E が新たに BMP の一部となった 79 文字のために設けた全般的な注意事項と矛盾するものである。これらの文字のすべては 2 バイト領域に見られる。したがって、その Unicode 符号位置は、4 バイト領域に対応付けられるはずがない。このマッピングは誤って追加されたものであると思われる。

**UPDATE:** 0x8139A634 から U+303E へのマッピングは、現在では訂正済みである。

GB 18030 の 4 バイト領域用の符号位置の数 (41,388——以下の所見を参照) から、論理的な帰結として、Unicode との対応がどこかで失われていると想定される。そして実際、GB 18030 は U+0080 への対応を提供していないのである。考えられる説明は、以下のようなものとなる。すなわち 1 バイト領域は、0x80 (「ユーロ」記号) を含むよう拡張された。それは正しく 0x80 の包含であり、この拡張は GB 18030 を GB 11383/ISO 4873 の適用範囲にとどめるものである。そして同時に、この符号位置は Unicode に対して非互換となる。GB 18030 の作成者は、次のように考えたわけである。U+20AC (「ユー

(7) U+20ACが0x80に対応するということは、規格票本文 [7.1] に明記されている（本稿7頁の6-aを参照）ように思われる。

ロ」記号）へのマッピングを4バイト領域で提供しないことで、暗黙のうちに0x80との対応を示そう、との。そうであるなら、UnicodeのU+0080への対応付けは、最初のマッピング登録として、符号位置0x81308130に割り当てられていなければならない。回答を与えられるべき疑問が残る。すなわち、この「穴」を埋めるために、U+0080から4バイト領域への対応付けを作成すべきなのだろうか。「自然な」解決としては、U+0080を、U+FFFFに対応する符号位置である0x8432EB38の後の、最初の可能な4バイト領域に割り当てる方法が考えられる。このやり方なら、4バイト領域の完全な再マッピングを回避することができるであろう（U+0080は、4バイト領域の最初の文字、0x81308130と対応を持つのが筋であったのだが）。これまで述べてきた点から、次のような一般的所見を付け加えさせていただきたい——GB 18030の1バイト領域からUnicodeへの明示的なマッピングがもしあれば、有用であろう、と。

**UPDATE:** 符号位置0x80（「ユーロ」記号が割り当てられていた）は、GB 18030の1バイト領域から削除された。予想できたことだが、Unicodeで未定義のU+0080は、現在では0x81308130からはじまるGB 18030の4バイト領域の最初の符号位置に割り当てられている。もちろんこれは後続するマッピングに大きな影響を与え、それらはすべて並べ直されることとなった。

c) 附属書 E (pp. 294ff.)

先にb)で述べたような状況にもかかわらず、事実上、4バイト領域における全体の符号位置の数には変化はない（「1増1減」である）。BMPは65,536の符号位置を持つ。GB 18030の94頁から296頁の表は4バイト領域中に41,388のBMP符号位置を示しており、残りのBMP符号位置は、1バイト領域（129）および2バイト領域（23,940）にあって、合計65,457である。BMPを完全に網羅するには「欠けて」いるように思われる79の符号位置は、附属書Eに「隠れて」いる。附属書Eに掲げられている79文字は、「臨時」としても「最終的な符号」としても附属書D（4バイト領域）の対応表には含まれない。文字自体は2バイト領域5（0xA989から0xA995）に見られ、「臨時」符号に対応付けられており、附属書Eの記述と一致する。これが、BMPの符号位置の総数と、BMPを網羅するためにGB 18030で用いられている符号位置の実数の違いの説明となる。

所見c)およびd)もまた関連を有する議論であるが、話を明瞭にするために、2つの段落で別個に検討する。

d) GB 18030の図2 (p. 6)

この図は、GB 18030の1バイト符号空間（0x00から0x80）を示している。符号位置0x24を表現するのに、なぜ通貨記号「円」が用いられているのかは明らかでない。GB 18030のなかで、GB 11383は、特定の符号空間をどのように用いるか——符号のみならず、文字についても（文字レパートリを記述しているGB 18030の[5]を参照）——を定める引用規格の役割を担っている。しかし、原規格のGB 11383-1989では——少なくとも1989年版では——符号位置0x24に「円記号」を割り当てるべきであるという決定は、明らかになされて

いない。その代わりに、別の段落 (7.4.2) は、「円記号」または「通貨記号」が意図または文脈に応じて用いられるであろうと説明している (この符号位置を表現する潜在的文字として、「ドル記号」さえも言及されている)。

この「円記号」は「ドル記号」によって置き換えるべきであると思われる理由がある。GB 18030 には、U+FF04 に対応する「全角ドル記号」0xA17E、そして U+FFE5 に対応する「全角円記号」0xA3A4 が存在し、どちらも正しいように思われる。これに加えて、Unicode の「円記号」U+005A に対応付けられる GB 18030 の 0x81308435 が存在するが、一方、U+0024 「ドル記号」へのマッピングは発見することができない。4 バイト領域に見られるのは 1 バイト領域や 2 バイト領域に含まれていない文字のみのはずであるという前提に立つなら、GB 18030 の文脈では 0x24 には「ドル記号」が割り当てられていると解釈できるであろう。GB 18030 は、GB 11838 は ISO 4873:1986 と同一であると強調している。しかし ISO 4873:1986 は、1986 年以来変更されていない。この規格は現在では、ECMA 43 の最新版と同一である。ECMA 43 は——現行の版における符号空間 (G0) で——U+0024 の「ドル記号」を含め、US-ASCII と同一であると定義されている。

**UPDATE:** ドル記号が符号位置 0x24 を「我が家」とするだろうという推測は妥当である。

e) Unicode のサロゲート領域を符号化する必要はあるのか？

述べてきたように、GB 18030 は完全な Unicode 文字レパートリに対して 1, 2, 4 バイトの符号位置を提供する。これは、Unicode のサロゲート領域 (U+D800 から U+DFFF) をも GB 18030 の 0x8336C830 から 0x83389837 の 4 バイト符号範囲に含む。この試みが冗長であると思われる理由が、少なくとも 2 つ存在する。

第 1 に、GB 18030 は、サロゲート・ペアで達成しようとするものと同じ結果を出すのに十分な符号位置を提供している。

第 2 に、GB 18030 の 4 バイト符号を Unicode の上位および下位サロゲートと同じやり方でペアとして用いたなら、事実上、8 バイト文字符号となってしまうであろう。GB 18030 内部でさえ、サロゲート・ペアを用いた場合と等しい符号化結果は、4 バイトのみで実現することができる。さらに、仮に 8 バイト符号が文字を表現するために用いられたなら、規格自体の規定に反するようになる。本当に Unicode のサロゲート領域を GB 18030 の 4 バイト符号空間に置くべきなのかという疑問には、正当な根拠があるように思える。

**UPDATE:** 再発行されたマッピング・データによれば、今後の規格では Unicode のサロゲート領域へのマッピングが含まれることはないであろう。U+D800-U+DFFF へのマッピングは、GB 18030 の 4 バイト領域のどこにも見出すことができない。

バージョン履歴:

- 1.1: 最初の公開バージョン。
- 1.2: GB 18030 の符号空間を記述する段落中の訂正。
- 1.3: マイナーな修正。その他の所見セクションに e) を追加。
- 1.4: 更新バージョン。再発行されたマッピング・データにおける大きな変更について、記述を追加した。GB 18030-2000 の符号空間を説明する図を追加した<sup>8)</sup>。GBK2K という表記を GB 18030 に変更した。

(8) この日本語化 PDF ファイルでは、図とその説明は省略した。オリジナルを参照されたい。