

2010年までのデータの文字コード変換について

1. 文字コードの相違

「JAPAN/MARC 2009 フォーマット」(以下「2009 フォーマット」という)までと「JAPAN/MARC MARC21 フォーマット」(以下「MARC21 フォーマット」という)では、文字コード及び文字の取扱いに相違がある。

◆2009 フォーマットまで

①漢字(2バイト)モード: JIS X 0208: 1990 (以下「JIS コード」という)

②英数字(1バイト)モード(制御文字符号含む): EBCDIC コード

③漢字(2バイト)モードの制御文字符号: JIS X 0207

漢字モードで表記したデータ部分では、JIS コードを持たない文字は可能な限り意味上・字形上関連する JIS コードを持つ文字に置換えるが、関連する文字がない場合等、追加文字を設定する。

◆MARC21 フォーマット

すべて Unicode (UTF-8) (以下「Unicode」という)

Unicode を持たない文字は可能な限り意味上・字形上関連する Unicode を持つ文字に置換えるが、関連する文字がない場合等は「=」に置換え、追加文字は設定しない。

「追加文字コード一覧」において、2009 フォーマットまでに設定した追加文字コードについて、対応する Unicode を示す。

3. 文字の置換え方針について

2009 フォーマットまで字体を統一していた関係上、Unicode 上異なる符号位置であっても字体を統一している場合がある。

3-1. 「JIS X 0208:1997」における「6.6.3 漢字の字体の包摂基準」にあてはまるものの中には、Unicode では異なる符号位置である場合があるが、JIS コードを持つ文字に置換えている。

例1)	賴	→	賴
	大漢和検字番号 36861 Unicode 8CF4		大漢和検字番号 43529' Unicode 983C JIS コード 4D6A

また、「6.6.3 漢字の字体の包摂基準」にあてはまる字形に Unicode が付与されている場合は、同字とみなし、対応する Unicode を記載している。

例2)	葳	→	葳
	大漢和検字番号 31456		Unicode 8473

3-2. 「JIS X 0208:1978」から「JIS X 0208:1983」に改訂した際、字体の変更があったものは、Unicode では異なる符号位置である場合があるが、変更後の字体を統一して使用している。

例3) JIS コード 89A8	鷗	→	鷗
	大漢和検字番号 47268 Unicode 9DD7		大漢和検字番号 補巻 769 Unicode 9D0E

3-3. 1997年12月以前に作成したデータでは、異なるJISコード及びUnicodeを持っている場合でも、旧字体を新字体に置換えたり、通用字形を統一して使用している場合がある。

例4)	學	→	学
	大漢和検字番号 7033 Unicode 5B78 JIS コード 555C		大漢和検字番号 6974 Unicode 5B66 JIS コード 3358

例5)	龍	→	竜
	大漢和検字番号 48818 Unicode F9C4 JIS コード 4E36		大漢和検字番号 25751 Unicode 7ADC JIS コード 4E35

4. 付表「追加文字コード一覧表」について

4-1. 区分 文字の種類を以下の9種類に大別する。

4-1-1. 記号 Unicode の以下の範囲に含まれる文字、及びUnicodeがない記号。

- General Punctuation (一般句読点、U+2000-206F)
- Geometric Shapes (幾何学模様、U+25A0-25FF)
- Mathematical Operators (数学記号、U+2200-22FF)
- Control Pictures (制御機能用記号、U+2400-243F)
- Spacing Modifier Letters (前進を伴う修飾文字、U+02B0-02FF)
- Miscellaneous Symbols (その他の記号、U+2600-26FF)
- Miscellaneous Symbols and Arrows (その他の記号及び矢印、U+2B00-2BFF)
- IPA Extensions (IPA 拡張、U+0250-02AF)
- Supplemental Mathematical Operators (補助数学記号、U+2A00-2AFF)
- Letterlike Symbols (文字様記号、U+ 2100-214F)

- Arrows (矢印、U+2190-21FF)

4-1-2. ラテン文字 Unicode の以下の範囲に含まれる文字。これらの文字に Combining Diacritical Marks (U+ 0300-036F) を付加したのものも含む。

- Basic Latin (基本ラテン文字、U+0000-007F)
- Latin-1 Supplement (ラテン 1 補助、U+0080-00FF)
- Latin Extended-A (ラテン文字拡張 A、U+0100-017F)
- Latin Extended-B (ラテン文字拡張 B、U+0180-024F)
- Latin Extended Additional (ラテン文字拡張追加、U+1E00-1EFF)

4-1-3. ギリシア文字 Unicode の Greek and Coptic (ギリシア文字及びコプト文字、U+0370-03FF) 範囲に含まれる文字のうち、ギリシア文字。これらの文字に Combining Diacritical Marks (ダイアクリティカルマーク (合成可能)、U+ 0300-036F) を付加したのものも含む。

4-1-4. キリール文字 Unicode の Cyrillic (キリール文字、U+0400-04FF) の範囲に含まれる文字。

4-1-5. 囲み記号 文字を丸で囲んだ記号。英数字、片仮名、平仮名、漢字を囲んだものがある。

4-1-6. 平仮名 Unicode の Hiragana (平仮名、U+3040-309F) の範囲に含まれる文字。

4-1-7. 片仮名 Unicode の Katakana (片仮名、U+ 30A0-30FF) の範囲に含まれる文字。

4-1-8. ハングル Unicode の Hangul Syllables (ハングル音節文字、U+ AC00-D7AF) の範囲に含まれる文字。

4-1-9. 漢字 Unicode の以下の範囲に含まれる文字。及び、Unicode を持たない漢字。

- CJK Unified Ideographs (CJK 統合漢字、U+4E00-9FFF)
- CJK Compatibility Ideographs (CJK 互換用漢字、U+F900-FAFF)

以下の範囲に含まれる文字は、運用上 CJK 統合漢字・CJK 互換用漢字に置換えるか、GETA MARK (ゲタ文字”=”, U+3013) に置換える。

- CJK Extension-A (CJK 統合漢字拡張 A、U+3400-4DB5)
- CJK Extension-B (CJK 統合漢字拡張 B、U+ 20000-2A6D6)
- CJK Extension-C (JK 統合漢字拡張 C、U+2A700-2B73F)

4-2. 追加文字コード 「JIS X 0208:1990」の追加文字等の範囲において、2009 フォーマットまでに定めた区点コード。

4-3. Unicode 『Unicode 6.0.0』で定められている符号位置。置換え先 Unicode を持たない場合は、「なし」とし、「置換え字 Unicode」欄で置換え先の符号位置を指定する。

4-4. 大漢和検字番号 『大漢和辞典』(諸橋轍次著 修訂第2版 大修館書店)にて付与されている番号。「h***」は「補巻***番」を指す。

4-5. 置換え字 Unicode Unicode を持たないか、運用上使用しないものについて、置換え先の Unicode 符号位置を示す。具体的な置換え文字が指定できない場合は、3013 (GETAMARK) を指定する。

4-6. 文字の説明 追加文字及び置換え文字の『Unicode 6.0.0』で定められている名前、またはその文字の説明。

5. 区切り文字の置換え

2009 フォーマットまで区切り文字として使用していた追加文字は、以下のとおり置換える。

5-1. // (追加文字コード: 2231)

①責任表示と役割表示との区切りに使用している場合は、SPACE(U+0020)に置換える。

②個人名標目における姓と名との区切りに使用している場合は、COMMA(U+002C) 、SPACE(U+0020)の2文字に置換える。

③件名標目と細目との区切りに使用している場合は、IDEOGRAPHIC SPACE(U+3000)、FULLWIDTH HYPHEN-MINUS(U+FF0D)、FULLWIDTH HYPHEN-MINUS(U+FF0D)、IDEOGRAPHIC SPACE(U+3000)の4文字に置換える。

上記①②③以外は、文字として“//”を使用している場合とみなし、PARALLEL TO(U+2225)に置換える。

5-2. \$ の縦棒が2本になっている形

印刷カードにおいて、サブフィールド開始文字の表現として使用していたもの。DOLLER SIGN(U+0024)に置換える。

(収集・書誌調整課)