

ISO10646 UCS と漢字

宮澤 彰 (学術情報センター)

1993/05/22

1 UCS とは

ISO-IEC 10646 Universal Multiple-Octet Coded Character Set(UCS) は Part 1:Architecture and Basic Multilingual Plane が昨年承認され現在印刷段階である。図-1 にその構成、図-2 に DIS の一部から文字ブロックを示す。図-2 でわかるように極めて多様な文字がふくまれている。

2 UCS と統合化漢字レパートリー

2.1 CJK-JRG

UCS の I-zone に含まれている CJK unified ideographs は CJK-JRG(China, Japan, Korea - Joint Research Group) によって作成された統合化漢字レパートリーをそのまま入れたものである。CJK-JRG は日本、中国、韓国、TCA(台湾)、HKTF(香港)、UNICODE などが参加し、91 年 7 月～92 年 3 月までの間に 4 回開かれた会議で(現在も会議は続いているが)統合化漢字レパートリーを作成した。このレパートリーは日本、中国、韓国、台湾の標準漢字コードを基にして、これらのうち同じと考えられる文字を「統合化」した文字表である。

2.2 Unification rules

何が同じと考えられるかについて CJK-JRG ではわずかな差は無視して同じ形のものを統合するというルールを立てた。何がわずかな差であるかは細かい表を作ってルール化することにした。図-3 に統合化ルールの一部を示す。

2.3 Source code separation

各国の標準コードとの互換性を保つために CJK-JRG では Unification rule に優先して Source code separation の原則を立てた。これは各国の標準で別の文字としていたものは形による統合の対象外とするもので、各国コードから UCS に変換した後逆変換しても情報のロスがないことを保証している。

3 UCS をめぐる問題

3.1 枠組み

問題点を整理するための枠組みとしてスクリプトの世界、交換コードの世界、内部コードの世界に分けて議論することを提唱する。(図-4)

3.2 何が漢字か？

これは本来スクリプトの問題であり、言語学の問題である。字体や、字形の問題を合わせ考えたとき何にコードを与えるべきかはまだ、合意の得られる段階ではないというのは確かである。しかし一方、こういった問題が意識されるようになったのはJ I Sコードができ、機械処理されるようになったためでもある。言語学的にこの問題に答えを出そうと思ったらおそらく今世紀中には終わらない問題であり、文字コードの作成をそれまで延ばすわけにも行かない。

3.3 字形字体問題

漢字は一般に複数の異体字を持ち、一つの字体でも筆法の差などにより、複数の字形になる、と言う字種—字体—字形ツリー構造モデルは、同一字体が複数の文字に属するという比較的少数の例外を除けば良く説明できるモデルである(図-5、6)。この欠点をカバーするものとして、3次元モデル(XYZモデル)も提唱されている(図-7)。問題は実際の文字をこれらのモデルにあてはめようと言うときに、言語により、時代により、さらに個人により意見が異なる点である。UCSでは複数字形(字体)の並置によりコード化されている「文字」の字形のゆれの範囲をある程度示している。

3.4 外字問題

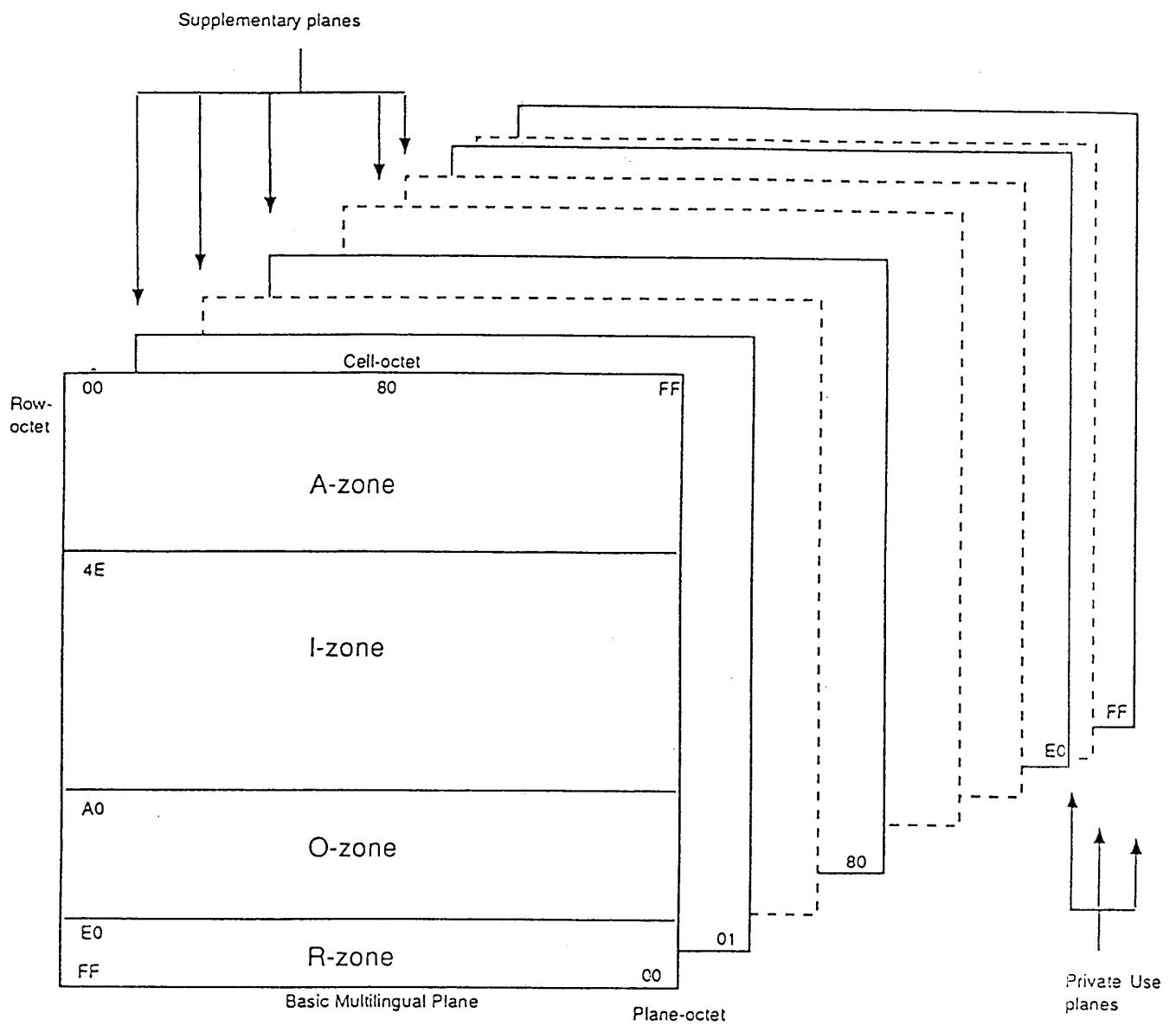
文字セットは(ある時点では)有限であるが、漢字はwordの世界と同じように閉じていない。現代においても、これが正しいかどうかは難しいところであるが、いずれにしても5、6万を越えてあることは確かなようであり、これまでに機械処理に乗った文字数ではまだコード化されていない文字(外字)が発生することは避けられない。この問題を避けるには登録制度(動的な文字セット管理)しかないが、その膨大さのために、技術的にも、制度的にもまだ、難しい段階ではないだろうか。

4 まとめ：何が解決し何が解決しないか？

UCSによって上に述べたようなスクリプトに関する問題点は解決しない。現在のJ I Sで起こっているのと同じく、これらの問題は継続する。ただし、外字は文字数の増えた分だけ発生確率が小さくなることは期待できる。(本質的な解決にはならないが)。UCSによって解決する問題の最大のものは同じ文字セットで複数の言語を扱えるようになるという点である。

また、現在のEUCとシフトJ I Sのような、同一交換コードに対する複数内部コードによる混乱という問題は、UCSのインプリメントにの仕方と運用によっては解決する可能性もある。

ISO/IEC DIS 10646-1.2



Labels A-zone, I-zone, O-zone, and R-zone are specified in clause 8.

図－1 UCS の構成

pecially applies to the digits, to the symbols, and to the use of Latin letters in dual-script applications.

Therefore, in using this coded character set, the reader is advised to refer first to the block names list in clause 19 or an overview of the BMP in figure 3, and then to turn to the specific code table rows for the relevant script and for symbols and digits.

19 Block names

The following list contains the blocks defined in the BMP. The block names are used in providing for the specification of subsets (see annex A for subset collections).

Block name	from	to
ISO-646 IRV	0020 - 007E	
LATIN-1 SUPPLEMENT	00A0 - 00FF	
EXTENDED LATIN-A	0100 - 017F	
EXTENDED LATIN-B	0180 - 024F	
IPA EXTENSIONS	0250 - 02AF	
SPACING MODIFIER LETTERS	02B0 - 02FF	
COMBINING DIACRITICAL MARKS	0300 - 036F	
GREEK	0370 - 03FF	
CYRILLIC	0400 - 04FF	
ARMENIAN	0530 - 058F	
HEBREW	0590 - 05FF	
ARABIC	0600 - 06FF	
DEVANAGARI	0900 - 097F	
BENGALI	0980 - 09FF	
GURMUKHI	0A00 - 0A7F	
GUJARATI	0A80 - 0AFF	
ORIYA	0B00 - 0B7F	
TAMIL	0B80 - 0BFF	
TELUGU	0C00 - 0C7F	
KANNADA	0C80 - 0CFF	
MALAYALAM	0D00 - 0D7F	
THAI	0E00 - 0E7F	
LAO	0E80 - 0EFF	
TIBETAN	1000 - 105F	
GEORGIAN	10A0 - 10FF	
ADDITIONAL EXTENDED LATIN	1E00 - 1EFF	
GREEK EXTENSIONS	1F00 - 1FFF	
GENERAL PUNCTUATION	2000 - 206F	
SUPERSCRIPS AND SUBSCRIPTS	2070 - 209F	
CURRENCY SYMBOLS	20A0 - 20CF	
COMBINING DIACRITICAL MARKS FOR SYMBOLS	20D0 - 20FF	
LETTERLIKE SYMBOLS	2100 - 214F	
NUMBER FORMS	2150 - 218F	
ARROWS	2190 - 21FF	
MATHEMATICAL OPERATORS	2200 - 22FF	
MISCELLANEOUS TECHNICAL	2300 - 23FF	
CONTROL PICTURES	2400 - 243F	
OPTICAL CHARACTER RECOGNITION	2440 - 245F	

ENCLOSED ALPHANUMERICS	2460 - 24FF
BOX DRAWING	2500 - 257F
BLOCK ELEMENTS	2580 - 259F
GEOMETRIC SHAPES	25A0 - 25FF
MISCELLANEOUS DINGBATS	2600 - 26FF
DINGBATS	2700 - 278F
CJK SYMBOLS AND PUNCTUATION	3000 - 303F
HIRAGANA	3040 - 309F
KATAKANA	30A0 - 30FF
BOPOMOFO	3100 - 312F
HANGUL JAMO	3130 - 318F
CJK MISCELLANEOUS	3190 - 319F
COMBINING HANGUL JAMO	31A0 - 31FF
ENCLOSED CJK LETTERS AND MONTHS	3200 - 32FF
CJK COMPATIBILITY WORDS AND HOURS	3300 - 337F
CJK COMPATIBILITY ABBREVIATIONS AND DAYS	3380 - 33FF
HANGUL	3400 - 3D2F
SUPPLEMENTARY HANGUL	3E00 - 458F
OLD HANGUL	4500 - 4C8F
CJK UNIFIED IDEOGRAPHS	4E00 - 9FFF
PRIVATE USE AREA	E000 - F7FF
CJK COMPATIBILITY IDEOGRAPHS	F900 - FAFD
ALPHABETIC PRESENTATION FORMS	FB00 - FB4F
ARABIC PRESENTATION FORMS-A	FC50 - F0FF
CJK COMPATIBILITY FORMS	FE30 - FE4F
SMALL FORM VARIANTS	FE50 - FE6F
ARABIC PRESENTATION FORMS-B	FE70 - FEFE
HALFWIDTH AND FULLWIDTH FORMS	FF00 - FFEF
SPECIALS	FFFD - FFFD

20 Characters in bi-directional context

A class of left/right handed pairs of characters have special significance in the context of bi-directional text. In this context the terms LEFT or RIGHT in the character name are also intended to imply "opening" or "closing" forms of character shape, rather than a strict left-hand or right-hand form. These characters are listed below.

0028	LEFT PARENTHESIS
0029	RIGHT PARENTHESIS
005B	LEFT SQUARE BRACKET
005D	RIGHT SQUARE BRACKET
007B	LEFT CURLY BRACKET
007D	RIGHT CURLY BRACKET
103C	TIBETAN LEFT BRACE
102F	TIBETAN RIGHT BRACE
207D	SUPERSCRIP LEFT PARENTHESIS
207E	SUPERSCRIP RIGHT PARENTHESIS
208D	SUBSCRIPT LEFT PARENTHESIS
208E	SUBSCRIPT RIGHT PARENTHESIS
2329	LEFT-POINTING ANGLE BRACKET
232A	RIGHT-POINTING ANGLE BRACKET
3008	LEFT ANGLE BRACKET
3009	RIGHT ANGLE BRACKET
300A	LEFT DOUBLE ANGLE BRACKET
300B	RIGHT DOUBLE ANGLE BRACKET
300C	LEFT CORNER BRACKET

図-2 UCSの文字ブロック表

5. Examples of differences of actual shapes(to be unified)

The examples shown here are considered to have the same abstract shape but have the different actual shapes.

讠·讠·讠, 示·示·示, 艮·艮·艮, 食·食·食,
 黃·黃, 盥·盥, 曷·曷, 包·包,
 青·青, 每·每, 冊·冊, 爭·爭,
 畵·畵·畵, 畵·畵, 步·步, 者·者,
 臭·臭, 并·并, 开·开·开, 骨·骨,
 呂·呂, 直·直, 渠·渠, 吳·吳·吳,
 眞·眞·眞, 爲·為, 單·单, 曾·曾·曾,
 成·成, 專·專, 内·内, 卅·卅

The differences are further classified as follows:

5.1 Examples of the difference in writing sequences

周·周, 告·告

5.2 Examples of the difference in rotated strokes/dots

半·半, 勺·勺, 羽·羽·羽,
 酋·酋, 兼·兼, 益·益

図-3 統合化ルール (一部)

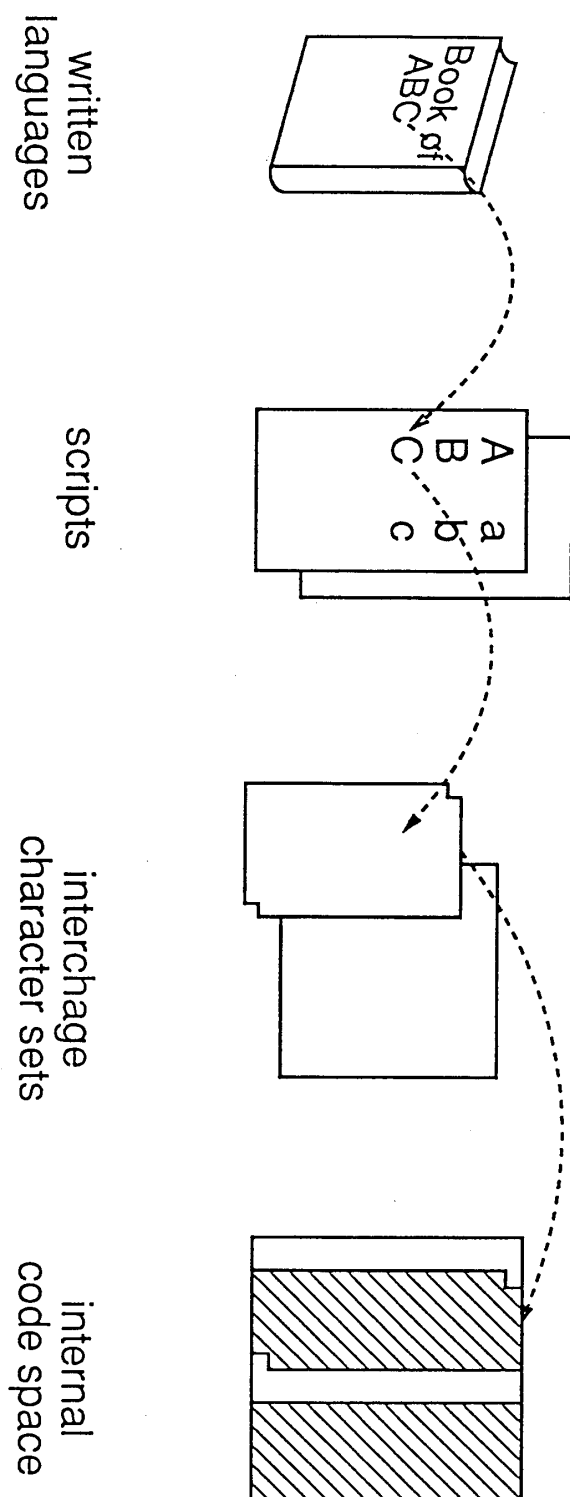


図-4 文字コード論の枠組み

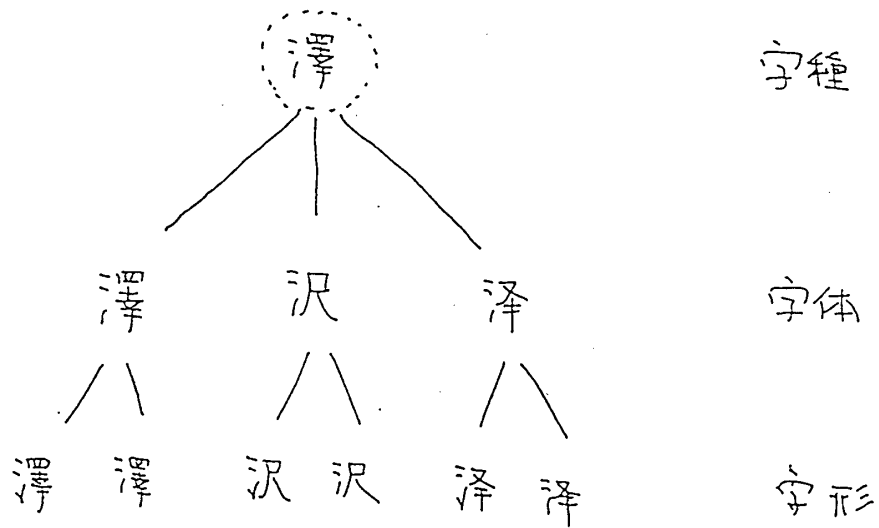


図-5 字種—字体—字形ツリー構造モデル

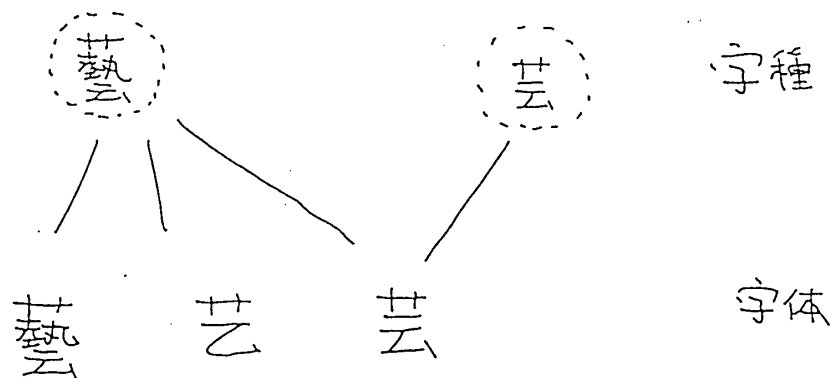


図-6 字種—字体—字形ツリー構造モデルの例外

Z (type face difference)

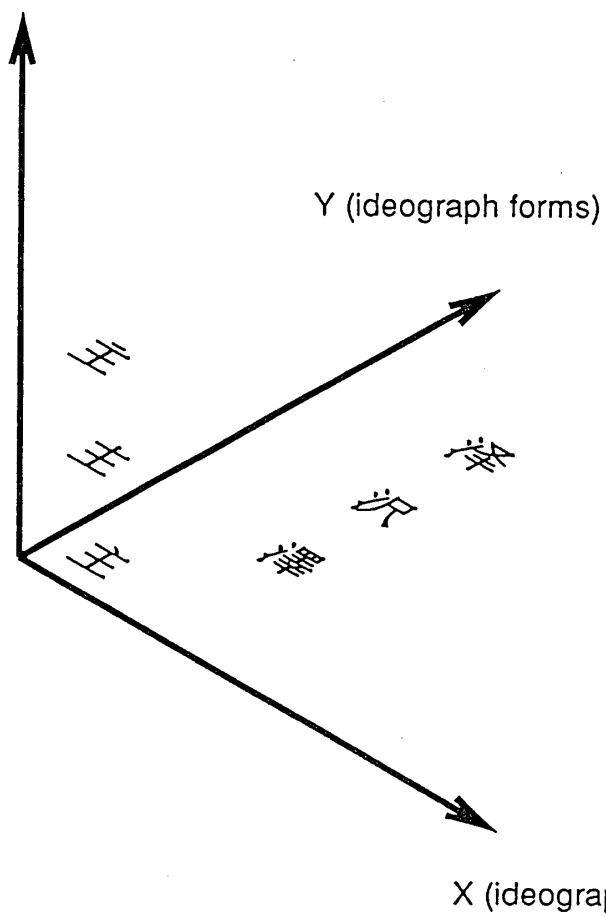


図-7 3次元モデル (XYZモデル)