電子化テキストの国際的共有

国際符号化文字集合(ISO/IEC 10646-1)のJIS化と今後の拡張

芝 野 耕 司

キーワード：符号化文字集合、文字コード、JISコード、漢字字書

要 旨

この論文では、まずこれまでのISO文字コード体系を検討する。その後、この検討を基に、国際符号化文字集合(ISO/IEC 10646-1：1933)の概要を紹介し、10646の位置づけを試みる。そして、JIS X 0221（案）の作成にあたっての問題点を検討し、今後の対応方針に関して述べる。

1. ISOコード

国際標準化機構(ISO)では、1967年に推奨仕様を出したISO R 646[1]以来、この主に7ビットコードを規定しているISO 646をいわばバイブルとして、順次、ISOコードの開発を行なってきた。ここでは、主要なISOコードの開発の歴史とそれぞれの開発の意図を分析し、国際符号化文字集合（ISO/IEC 10646-1）[2]へいたる道について述べる。

このISO 646は、情報処理と通信の両方の分野での文字コードの規格統一を図ると共に、それ以前に通信分野で用いられてきたCCITT Alphabet #2での5ビットコードを、よりコード空間の広い、6ビット又は7ビットに拡張することによって、より適切に文字を扱うことを可能にしようとしたものである。

このISO 646の設計には、次のような幾つかの原則が含まれる。

（1） 処理と交換用コード

（2） 図形文字のコード化

（3） 各国及び適用業務毎の版の作成

（4） 合成文字

まず第1の特徴は、新しく生まれた情報処理での文字コードと情報交換、すなわち、通信での文字コードとの統一を目指した点である。第2の特徴は、文字のコード化を文字の意味でも音でもなく、図形としての文字にコードを与えた点である。

また、5ビットから7ビットまで拡張されたとはいっても、依然として制限がきつく、足りないコード空間での処理及び交換を行なうために、図形文字に割り当てた94文字のうち、共通の82字以外を各国文字又はアプリケーション毎の文字用に割り振り、646をもとにした複数の版の開発・利用を原則とした。同時にこの12字では、ヨーロッパのアクセン

ト付きの文字の表現には足りないことから、BACKSPACE シーケンスによる合成文字を導入した。

ASC II コード、JIS 1 バイトコードなど各国のコード及び通信、書誌学分野などの世界のすべてのコードは、基本的には、この ISO 646 の版として開発されてきた。

アクセント付きの文字を必要としないアメリカにおいても、少なくとも三つ以上の版が必要である、とこの時期に ANSI から報告されている。[3]

ISO 646 での複数の版の利用には、コード拡張が必要になる。このコード拡張性を規定したのが、ISO 2022[4] である。

ISO 2022 の基本的な機能は、ISO 646 のそれぞれの版で異なる 12 字を 1 文字づつ呼び出し利用するシングルシフト、版全体を取り替えるロッキングシフトの二つである。

しかし、ISO 2022 は、文字コードの拡張を可能にする規格ではあっても、拡張を保障する規格ではない。事実、ISO 2022 の適合性では、共通するエスケープシーケンスをもとにした情報交換を行なうことを規定しており、このことは、2 バイト文字の取り扱いを事実上、拒否することになる。

ISO 646 と ISO 2022 からなるコード系に対して、ヨーロッパでは、次の問題点が次第に明らかになってきた。次の項目は、Latin 1 として広く知られる西ヨーロッパで現在用いられている 8 ビットコードを含む規格群である ISO 8859[5] の DIS 投票時に添付された 8859 略史からの引用である。

— これらの版は、必要とされるであろうすべての図形文字を提供しない。
— 幾つかの文字、特に、アクセント付きの文字については、BACKSPACE シーケンスに頼る必要があり、このような合成文字を含むデータ処理を行なう際に問題をうむ。
— 異なった版間での情報交換は、実際上、(94 字ではなく) 82 の共通図形文字に限られる。

これらの問題を解決するために、8859 が次の設計思想で設計された。次の項目も、ISO 8859 の DIS 投票時に添付された 8859 略史からの引用である。

— このコードは、特定の地域のすべてのユーザに対して同じものであり、
— すべての図形文字の単一バイトコード化を提供し、簡単に処理できるようにし、
— 産業界で用いられている文字集合を考慮に入れる。

しかし、この文字コードも、次第にその限界を示すようになった。特に、統合ヨーロッパを目指す EC 内での文字コードの利用を考えるといらなくしたはずの ISO 2022 が再び必要になると共に、コード拡張が必要でない一つのコード表を構成する言語の選択がきわめて難しくなってきた。

すなわち、8 ビットは、ヨーロッパにとって、どのように捉えても、十分では無くなったのである。一方、依然として 7 ビットの ASC II コードを利用してきたアメリカにおいても、従来の数値計算及びデータ処理から、文書処理、DTP へとコンピュータのアプリケーション分野の広がりによって、7 ビットでも、8 ビットでもコード空間が足りなくなってきた。

そして、前述の8859と同様の設計思想で開発されたのが10646である。

2. 国際符号化文字集合(ISO/IEC 10646-1)の概要[2]

国際符号化文字集合(Universal Multiple-Octet Coded Character Set: UCS, ISO/IEC 10646-1)は、世界の言語の記述した形(記述形)及び記号の表現、伝送、交換、処理、蓄積、入力及び表示に適用できる16ビット又は32ビット単位の新しい文字コードの国際規格である。

この規格は、UCS-4と呼ぶUCSの4オクテット(32ビット)正規形とUCS-2と呼ぶUCSの2オクテット(16ビット)BMP形を想定している。ここで、BMPは、基本多言語面(Basic Multilingual Plane)の頭字語である。

また、UCSの基本多言語面(BMP)を規定し、世界の言語の用字及び記述形で使用する図形文字の集合を定義している。図1に10646の構造を示す。

そして、この規格は、全く新しい文字コードに関する国際規格であり、10646という規格番号が示唆するように、これまでバイブルであった646の後継にあたる。事実、今後順次、この規格中で規定している図形文字の名称に他の文字コードの規格を合わせることになっている。

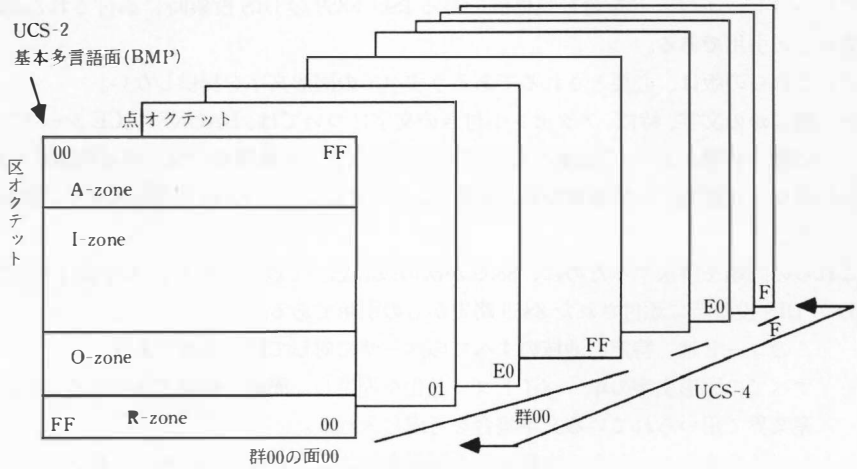

図1　10646の構造

UCS-4では、各文字は群(Group)，面(Plane)、区(Row)及び点(Cell)の4バイトで表す。UCS-4の群00面00は、特に基本多言語面(BMP)と呼び、群及び面を省略することができる。この省略形がUCS-2である。

基本多言語面には、これまで世界中で文字コードとして用いられてきたほとんどの図形文字、34,000字強が収録されている(表1参照)．すなわち，16k文字の連続領域として保留されているO-zoneを含め、基本多言語面には、約30,000字の空き領域がある。

A-zoneは、アルファベットを中心に漢字(CJK統合表意文字)以外のほとんどの用字及

び記号を含む。この A-zone には、韓国のハングルも含むため、日中韓を中心とする漢字文化圏以外の文字は、記号類を含め、5,000 字強にすぎない。

表 1　基本多言語面の各領域文字数

コードポイント		文字数	CP 占拠率	文字占拠率
A-zone	19903	11764	30%	35%
I-zone	20992	20902	32%	61%
O-one	16384		25%	
R-zone	8190	1375	13%	4%
合計	65469	34041		52%
ハングルを除く A-zone	5142		15%	
Hanglu		6622		
記号類		1377		

I-zone は、約 21,000 字の漢字（CJK 競合表意文字）のための領域である。

O-zone は、将来の拡張のために確保している領域である。そして、R-zone は、互換文字などを含む制限利用領域である。

10646 は、646 の版と同様の合成文字を含むコードである。しかし、このコードでは、合成文字に関する大幅な明確化と実装水準分けを行なっている。この意味で、10646 は、646 の合成文字の概念の改良を行なっている。

もう一つの大きな変更点は、この規格で規定する図形文字集合のサブセットを容認あるいは推奨している点である。この規格では、各文字を数え挙げ型で指定する制限部分集合と、あらかじめ決められた部分集合を用いて指定する選択部分集合との、二つの指定方法を用いて、サブセット化を行なうことができる。

こうした文脈で、この規格の本来の意味を考えると、次のように捉えることができる。すなわち、この規格では、世界中の図形文字に対して、一意にコード表現を与えることによって、世界中のどこでも、図形文字を交換可能にすることを最大の目的としている。もしかりに、この規格の目的が同時多言語処理であれば、サブセット化の容認は、この目的に反する。

また、従来の ISO 2022 が情報交換当事者間での明示的な取り決めを必要としていたことに対して、この規格では、こうした取り決めを必要としないことを前提としている。すなわち、この規格自体にすべての図形文字を含むことによって、この規格自体が取り決めとなる。

. この 2022 との違いは、適用分野の違いとしても理解することができる。2022 では、磁気テープによる情報交換などのように、特定に相手との間で情報交換をすることを前提としていたことに対して、この新しい規格では、インタネットのように、情報交換の相手を特定できないような新しいネットワーク環境での利用も、想定している。

3. 国際符号化文字集合のJIS化方針

JISにおいては、今後、この国際符号化文字集合への移行を促進することを検討している。これは、ソフトウェアの国際化の促進と国内のJISコード問題の解決には、この新しいコード系に移行すること以外にないと思われるからである。

3.1 ソフトウェアの国際化と日本の対応(SC 22ガイドライン)

ごく最近まで国際規格では、日本語を含む2バイト圏の言語を適切に取り扱うことができなかった。プログラム言語などの情報処理の国際規格中で各国文字集合を適切に取り扱うことへの最初のISOとしてのコミットメントは、1985年に国際規格原案(DIS)を出したISO/IEC 9075データベース言語SQLである。

こうした状況を改善するために、日本は、日本語機能専門委員会を1987年に設け、日本語だけではなく、広く非ラテンアルファベット圏、特に2バイト圏の文字を国際規格で適切に扱うことができるようにISOに求めるとともに、それぞれの技術分野で仕様提案を行なってきた。

特に、プログラム言語規格を担当しているSC 22では、日本提案を受けて、1987年に各国文字集合リポートに対するガイドラインを採択し、すべてのプログラム言語で2バイト文字の適切な取り扱いを求めた。[6]

このガイドラインでは、すべてのプログラム言語に対して、データ型(data type)、リテラル(literal)及びコメント(comment)として2バイト及び1バイトの各国文字を取り扱うことを可能にすることを必須項目として求めた。[6]

また、識別子として2バイト文字が使えるようにすることも要請した。

10646はこうした努力に対する全面的な解決策にはならないが、大きく貢献することが期待される。ソフトウェアの国際化に対する日本の努力との整合性を求め、今後日本は、10646の推進に努めるべきである。

3.2 JISコードの問題点

まず、最初の漢字符号、JIS C 6226-1978(JIS 78)[7]とは何であったのであろうか。この符号系には、通常の日本語の用字(script)では用いないギリシャ文字及びロシア文字を含むと共に、合成文字を含めるとすべてのヨーロッパ系の文字含んでいる。この点を考慮すると、JIS C 6226の設計意図は、その適用範囲にあるように“通常の国語の表記”とは見れず、10646と同じ自己完結的な国際符号(universal code)を目指したものと理解せざるを得ない。

このように理解しない場合、このコードは、ISO 646の原則、すなわち、同じ図形文字にはただ一つのコードポイントを与えること及びISO 646中の共通文字82文字のコード位置を変更してはならないという点に反することになる。

しかし、実際には、片仮名を含むISO 646の版として位置付けられるJIS C 6220-1969[8]

と併存して用いられることとなった。このことが2重符号化（duplicate coding）を含む多くの問題を惹起する。

結果として、国内の現状は、符号化文字集合の二つの要素、すなわち、文字集合及びビット組合せを決める規則の両方の面で実効的な標準が存在しない。

また、文字表中に保留領域を設けなかった結果、既存のJISコードを保守していくことが不可能になった。事実、最も広く用いられているパソコンでは、いわゆるJIS 78をもとにした自社拡張コードを用い、83年改正及び90年改正に全く追従していない。90年に制定した補助漢字に関しては、実質上、全く普及しなかったといってよいであろう。

すなわち、JISコードは、漢字に関しては、実質的には、標準にならなかった上に、ガイドラインとしても、その地位を失ってしまったといわざるを得ない。

3.3 国際符号化文字集合のJIS化

JIS化にあたっては、次の四つの附属書をつけることを予定している。

附属書1（規定）日本文字サブレパートリ

附属書2（規定）日本語データの情報交換用としての私用文字の使用制限

附属書3（参考）JIS X 0201-1973, JIS X 0208-1990及びJIS X 0212-1990との対応表

附属書1は、日本語文書で用いる文字集合を明確化する日本文字サブレパートリについての規定である。附属書2は、国内でのメーカの独自文字拡張を禁止することを意図している。附属書3は、既存のJISコードとの対応表である。

こうした幾つかの附属書を付けることによって、混乱を最小にし、10646の普及促進を図ろうとしている。

このJIS原案は、1994年3月にJIS X 0221(案)として、作成、1994年9月に工業標準調査会情報部会で審議され、1995年初頭にJISとして制定される見込みである。

4. ユーザから見た10646の利点

最大の利点は、ソフトウェア産業の自由化にあろう。世界中のソフトウェアが即時に、そして漢字化の費用をかけずに安価に購入でき、日本で作られたソフトウェアがファミコンソフトのように世界に出て行くきっかけとなることを期待している。

一方、ユーザから見た10646の最大の利点は、同時に多くの文字が利用可能になることである。そして、十分なコード空間を確保していることから、今後永久にこのコードを用いていくことが可能になると思われる。

また、前述のように実際に表示可能な文字種に関しては、サブセット化を許していることから、これまで世界中の表示装置で表示可能にするためのコスト面の問題から、追加することが実質上不可能であった多くの文字の追加が可能になった。

統合漢字コードも大きな成果であると考える。この統合漢字コードの実現には、90年JIS改正委員会より継続的に文字フォント普及センターで続けてきた、異体字の取り扱いに関する研究成果が統合基準の決定に支配的な影響を与えた。

この漢字コードの統合によって、従来個別に作成されてきた日中韓の情報流通が促進され、学問分野を含む文化交流が促進されることが期待される。

IPA 記号及び多くの記号類が追加されたことも、国語学及び言語学分野では大きな救いになると思われる。

5. 漢字表とその適用範囲

文字コードのための文字集合の設計に対して、常用漢字表、康熙字典又は大漢和辞典などの漢字にコードポイントを割り振れば済むといった乱暴な議論が一部にある。しかし、この議論自体は、これまでも間違いであったし、これからも間違いである。すなわち、どのような文字表であっても、その文字表自体の適用分野が存在し、その適用分野と漢字コード規格の適用分野が必ずしも同一ではないことによる。

例えば、常用漢字表では明示的に固有名詞を適用除外としている。常用漢字表だけにコードポイントを与えると仮定すると、この系では、埼玉も、岡山も、大阪も表記できない。

漢和辞典では、主として四書五経などの中国古典を読むことを主な利用分野としており、伝統的には国字の採用にはきわめて冷淡な態度を取ってきている。データ処理分野での利用を考慮すると、住所氏名で利用される国字を勢力的に収集しなければならない。

また、最近での広い分野での利用を考慮すると、国文学、国史学、東洋学、仏教学などの分野での漢字の収集が必要となる。

文字コードが含む文字集合の大きさに関しても、意味をもつことになる。

2000 字水準の文字集合は、日本の常用漢字表、中国の簡化字総表の水準にあたり、この水準での統合漢字表を作成しようとすると日本の常用漢字表中の通用字体と中国の簡化字の統合が必要となる。すなわち、この水準では、国語政策の問題となるのである。

また、この水準の文字では、前述のように国内では県名すら満足に表記できない。

JIS 漢字などの漢字コードの第 1 水準に当たる 3000 字水準では、基本的な日常の文書の表記が可能になり、日常使われる漢字コード水準である 6000 字あれば、表記の上で、日中ともほとんど困ることがないことが知られている。国語辞典などに集録されている漢字もこの水準である。

一般の印刷会社及び新聞社などで常時用意している漢字の水準が 10000 字水準である。通常の漢和辞典に収録されている漢字数は、この水準である。

10646 の収録水準である 21000 字に対応するのが、国内では 1 冊本としては最高の漢字数を誇り、その印刷には、10000 字以上の活字を新たに作ることが必要であった漢和辞典の漢字数に対応する。

そして、50000 字が最大規模の漢字字書である康熙字典及び大漢和辞典の漢字数である。

このように 10646 の統合表意文字の開発は、現代のコンピュータ時代の字体レベルでの漢字字書の作成の試みであると言える。この試みに対して、2000 字レベルの文字は当然含まれるが、このレベルでの議論がなされることは全くの間違いである。

また、一つの漢字は、多義的であり、各国での代表的と称する意味の一つを採り違いを論じることは全くの間違いである。このような間違った例には、“湯”の意味の違いを論じることや、“机”と“機”の違いを論じることがあるが、このような論を述べられる方は、一度、お持ちならば、手元の漢和辞典を引くことをお勧めする。

表 2　文字集合の大きさとそのもつ意味

2000 字	政策漢字—常用漢字表(1945 字),簡化字総表(2238 字)
3000 字	日常文書水準
	固有名詞の追加が必要
6000 字	日常文字コード，国語辞典
10000 字	漢和辞典，一般印刷可能水準，総正字数
21000 字	大規模辞書
50000 字	最大規模辞書

6. ISO IRG と今後の方向性

新しく創設した ISO/IEC JIC 1/SC 2/WG 2 IRG（Ideographic Rapporteur Group）で検討する統合表意文字には、幾つもの問題がある。最後に、IRG の課題を整理し、今後の方向性についての提言を述べよう。

IRG の最大の課題は、漢字の同定の問題である。用字表を前提とする名称を用いたこれまでの ISO 文字コードでの文字の同定の方法も、実質的には 10646 での記号類の同定の面では、既に破綻をきたしている。しかし、統合表意文字に関しては、まさにこの問題を原規格に委ねただけで、全く規格本文中には何もない。少なくともこの問題を解決する必要がある。

この課題に対しては、CJK-JRG で作業用に作成した既存字書との対応表を取り込むことがまず最初の手段として考えられる。もう一つの方法は、異体字対応表を導入することである。日中韓でこの異体字対応表についての合意を形成することは現行字体をもとにすると無理であると思われる(理由は、国語政策による文字の簡化にある)。しかし、康熙字典の正字を正字とし、これに対する異体関係を日中韓で統一し、各国毎の異体関係表は、康熙字典の正字と各国の政策文字との対応によって決めてはどうか。

次に問題は、並び順である。これからの追加では、すでにある統合表意文字の中間に文字を追加することはできない。すると、並び順は、別に照合順番表（collation table）などの分離した形で、対応する以外に方法はない。

すなわち、ここでは、文字クラスと照合順番の分離を行なわなければならなくなる。

最後の問題は、ソースセパレーションの問題である。前述のようにこの原則に従って分離した字形は、基本的には既存文字コードとの互換性のための字形であり、本来は、互換領域に位置付けるべきものであった。この割り当て位置の変更はできないし、すべきではないが、これらの文字がどの文字であったかについては、早急に一覧表を作り、10646 の

defect として、処理すべきであると考える。加えて、各国で作成するサブレパートリ中には、これらの文字を入れないようにする必要がある。

参考文献

[1] ISO R646: 1967, 6 and 7 Bit Coded Character Sets for Information Processing Interchange.
[2] ISO/IEC 10646-1: 1993, Information technology-Universal Multiple-Octet Coded Character Set (UCS)-Part 1: Architechture and Basic Multilingual Plane.
[3] 情報処理学会規格委員会編, 電子計算機の国際標準化, 情報処理学会, 1971.
[4] ISO 2022-1986, Information processing —ISO 7-bits 8-bits coded character sets— Code extension techniques.
[5] ISO 8859-1: 1987, Information Processing —8-bit single-byte coded graphic character sets— Part 1: Latin alphabet No. 1.
[6] Japan, Requirement for Character Handling in Programming Languages, ISO/IEC JTC 1/SC 22 N 460, Sept., 1987.
[7] JIS C6226-1978, 情報交換用漢字符号, 日本規格協会
[8] JIS C 6220-1969, 情報交換用符号, 日本規格協会

——東京国際大学助教授——