

文字図形識別番号のJIS化への取り組み

インデックスフォント研究会

板倉和治 黒田信二郎 長村玄

1. はじめに

10万字を越える文字図形に一意的な番号(文字図形識別番号)を与えた標準を定め、その文字図形識別番号によって当該文字図形の指定を行えるようにすることを目的とする標準のJIS規格原案(以下、本規格原案)の作成作業が現在進められている。

出版印刷の現場では顧客の要求により異体字等を正確な文字図形で表現することが必要になることが多くあり、要求された文字図形を印刷会社内部の文字コードと対応させたい一方で、そこに無い文字についてはいわゆる外字処理で対応している。外字処理は印刷会社に渡すテキストの作成者や、印刷会社までの流れの過程でも存在し、外字処理とその周辺で発生する様々な作業は日本語文化のコンピュータ処理の上での効率を悪化させる要因になっている。本規格原案はこの問題を解決もしくは軽減することに寄与するものである。

本規格原案作成の作業は非営利活動法人 文字文化協會の呼びかけで発足した「文字図形標準化検討委員会」で行われており、今年度中の作成を目指している。私共インデックスフォント研究会のメンバーもこれに参加して、規格原案の作成に中心的な役割を果たしている。本稿ではその規格原案について、その目的、必要性、内容等について記述する。

2. 必要性の背景

コンピュータで文字を扱うさまざまな業務の中で、異体字等を区別して印刷・表示したいという需要は極めて多い。特に人名などの固有名詞、学術文献・歴史資料などを扱う用途では、異体字等を区別できることがシステムの必須要件である。

ISO/IEC及びJISなどの情報交換用文字符号(以下、“文字コード”という)の収容文字数は、改訂もしくは新しい規格が制定されるたびに拡大されてきたが、上記の要求に応えるものではない。これは単に収容文字数の問題ではなく、文字コードは建前として文字図形を定義するものではなく、文字の種類を定義するものとされているからである。言い換えると、文字コードにおいては、包摂を前提としているため、包摂されている多くの異体字は、人名表記などで区別して指定したくても、それができない。(包摂とは同じ文字の複数の異体字に同一の符号を振ることである。)

また、文字コード収容文字数が拡大してきているといっても、冒頭の目的の用途の中にはそれでもまだ足りない文字がある場合がある。文字コードの仕組みの中では、そのような文字コードの未定義文字については、定義(符号化)されている文字と同様な簡便な方法による指定ができない。

3. 外字の存在と問題点

出版印刷分野では指定された文字図形を忠実に表現することが求められ、その文字図形が、文字コード規格の包摂文字に該当する場合や、そもそも該当する文字が定義されていない場合には、規格のコードを使わず外字を利用して処理を行っている。

印刷会社の内部では要求された文字図形を印刷会社内部の文字コードと対応させたうえで、そこに無い文字についてはいわゆる外字処理で対応している。しかも、多くの場合、印刷物ごとに外字の管理をしているのが現状であり、印刷データは外字を含むが故に共有性がない。

また、出版印刷業界では業務の電子化が進んでおり、同時に、コンテンツの電子化、電子書籍化への需要も増加している。同業界に限らず紙を直接の媒体にしないデジタルアーカイブシステム及びインターネットを利用するさまざまなアプリケーションにおいても、外字の必要性が存在するが、文字コードによる外字処理は事実上不可能である。代わりに外字を画像形式の情報で表現することがある。その場合は、その画像情報の生成段階で原稿で指定される文字図形の特定を行って画像を作成するという外字処理が行われる。

外字と言うやっかいな存在は出版印刷分野に限らず、企業、学校及び行政での人名を扱う業務などにおいても同様である。これらの業務では、文字図形同定作業に手間がかかるだけでなく、本来の業務外である、外字作成・管理、及びそれらのための事務処理なども大きな負荷になっている。

外字処理には、特定の文字図形が定義された外字コードを用いて文章中で外字を表現する方法や、外字を画像形式で表現したものを文章中に入れ込む方法などがある。そのことに関する処理自体が負荷となることは自明であるが、外字問題の本質は、外字はローカルなものであり、外字として表現されたものがどのような文字図形を示しているのかを理解することは、それが扱われる広義のシステム環境に依存するため、システム間で外字を含む情報の伝達には大きな困難を伴うという点にある。

4. 文字図形識別番号が果たす役割

本規格原案では、文字図形に一意的な番号（文字図形識別番号）を与えた標準を定め、その文字図形識別番号によって必要な文字図形の指定を行える。これにより、包摂のために区別できない文字図形や文字コード規格未定義文字を扱う場合、文字に習熟した特定の技能や、文字図形を特定するためのシステム機能の必要性が軽減される。

文字図形識別番号は、このように文字コード規格における包摂及び文字セットの制限によって需要を満たすことができない部分を補完するとともに、出版印刷業界をはじめ、関連システム業界及び行政サービスの窓口業務などにおいて、当該規格を文字図形の確認のための汎用的な基盤とすることで、煩雑な文字同定作業及び個別の外字フォントセットの管理業務の効率化が図れるほか、個別に規定されたさまざまな文字番号（文字コード）間の共通な参照番号としての役割を担うこともできる。

5. 文字コードJISとの役割分担

先に、文字コードJISでは「文字コードは文字図形を定義するものではなく文字の種類を定義するものとされている」と書いたが、実社会においては必ずしもそのように理解され、利用されているわけではない。実際、我々が認識するのは文字コードではなく、目に見える文字図形そのものである。システムによって表示される文字図形が異なっていると、不都合が生じる場合がある。文字コードJIS自身も「文字の種類を規定している」と言いながら、2004年の改訂で、表外漢字字体表の印刷標準字体に従って、規格票例示字形の微細な変更を行っている。

このようなことからシステムベンダーやフォントメーカーは、文字コードJISに製品が合致する為には、例示字形と一致していなければいけないと認識して行動している。そのため、常に文字が足りないと言う不満がJISに対してあり、実際に改版の度に収容文字数は拡大されてきた。また、拡大された新たな文字コードの中には既存の文字コードの異体字が例示字形として示されているものがあったり、既存の文字コードの例示字形が変更になるケースもあり、これらが更に混乱に拍車をかける原因ともなっている。

文字コードは文字の種類を表現するものであり、それが示す文字図形にバリエーションがあることは、膨大な数の表意文字で情報交換を行う為には必然の考えである。その考えからすれば、資源に限りのある文字コードとしては、微細な字形差に違うコードを振るべきではない。またそのことは文字コード本来の目的である情報交換を効率の悪いものにするということでもある。しかし、本規格原案のように、文字コードとは別の次元で文字図形を示すリファレンスがあれば、文字図形を正確に指定する必要性のあるときにのみ、それを使うことで問題は解決するか、あるいは混乱を小さくすることが出来る。

現在進められている本規格原案は、これまで説明してきたことから理解できるように、既存の文字コードJISとは全く異なる概念の規格である。

6. 本規格原案の内容

規格名称

識別番号を持った印刷参照用文字図形の集合（仮称）

適用範囲

この規格は、印刷及び表示における運用上区別して使用される文字図形の集合、及び文字図形の形状を識別するための識別番号を規定する。

定義

この規格で用いる主な用語及び定義は、次による。

文字図形：

印刷及び表示における運用上区別して使用される文字の形状。

文字図形識別番号：

文字図形に対して一意に振られた、000001 から始まる 10 進数六桁からなる番号。

識別番号を持った印刷用文字の図形集合

印刷用文字図形の集合として、10 万個の文字図形を規定する。(詳細については検討中)

印刷用文字図形は、いわゆる漢字の文字図形とする。

当該識別番号は 10 進数 6 桁のすべての桁を用いて表記する。

印刷用文字図形と対応する当該識別番号の関係を、付属書 A (規定) に印刷用文字図形一覧として示す。

付属書 A

付属書 A は定義された文字図形が文字図形識別番号順に収録されたものである。

これとは別に参考として文字を構成する部品の画数順に整理された付属書 B が存在する。

図 1 付属書 A の一部

一	丁	丂	丂	古	七	上	下	方	万	丈	三
000001	000002	000003	000004	000005	000006	000007	000008	000009	000010	000011	000012
上	下	丌	且	丂	丂	不	与	丂	丂	丑	且
000013	000014	000015	000016	000017	000018	000019	000020	000021	000022	000023	000024
丘	丈	丘	刃	且	丂	世	世	丘	止	丙	引
000025	000026	000027	000028	000029	000030	000031	000032	000033	000034	000035	000036
亥	斗	丙	丞	丟	兂	丽	系	北	兩	此	𠂇
000037	000038	000039	000040	000041	000042	000043	000044	000045	000046	000047	000048
兂	州	兩	𠂇	並	並	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇
000049	000050	000051	000052	000053	000054	000055	000056	000057	000058	000059	000060
𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇
000061	000062	000063	000064	000065	000066	000067	000068	000070	000071	000072	
中	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇
000073	000074	000075	000076	000077	000078	000079	000080	000081	000082	000083	000084
𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇
000085	000086	000087	000088	000089	000090	000091	000092	000093	000094	000095	000096
𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇
000097	000098	000099	000100	000101	000102	000103	000104	000105	000106	000107	000108
𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇	𠂇
000109	000110	000111	000112	000113	000114	000115	000116	000117	000118	000119	000120

7. 文字図形番号の運用事例

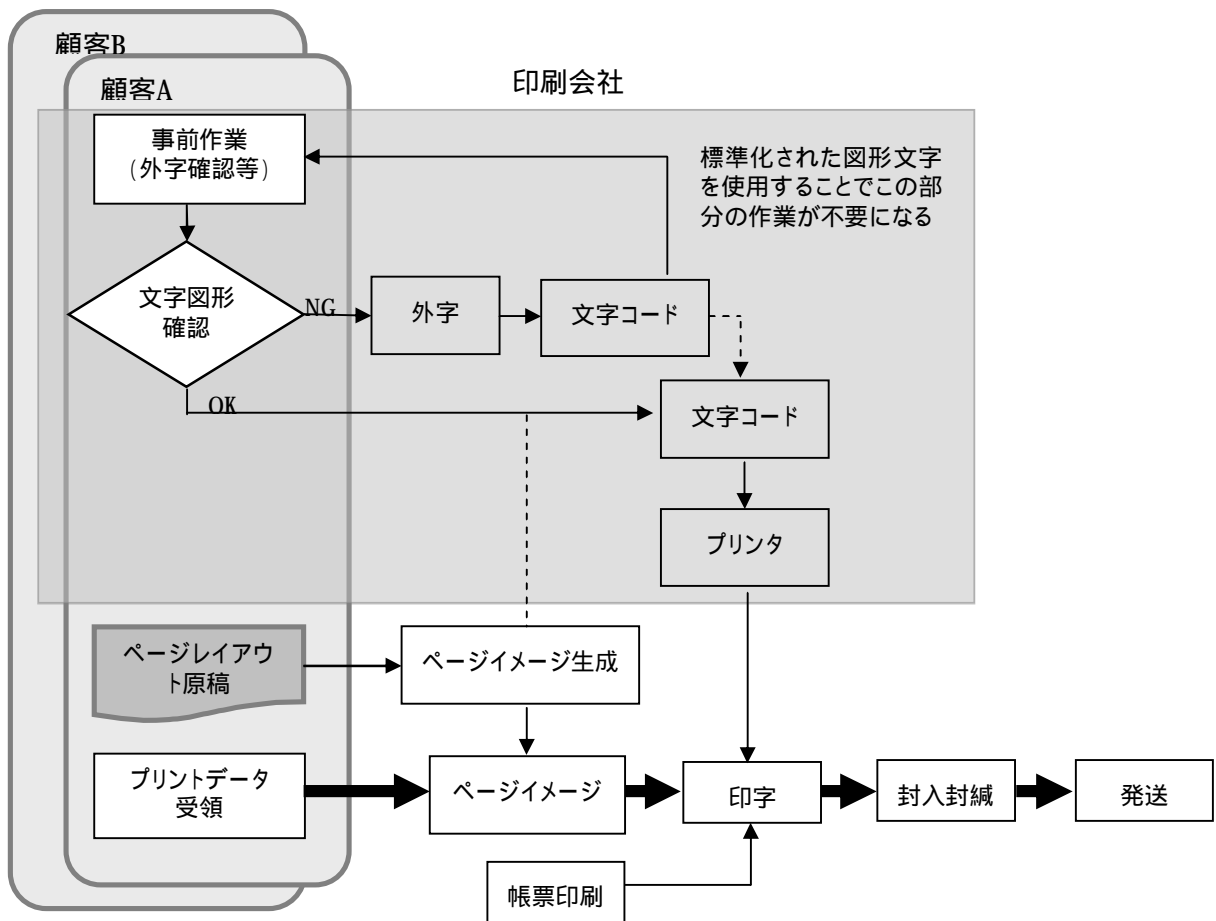
データプリントサービス

データプリントサービスは、印刷会社が顧客からクレジットカードの利用明細書、携帯電話の

利用明細書、ダイレクトメール、保険料控除証明書、及び各種通知書などのプリントデータを受取り、所定の帳票にプロダクションプリンタと呼ばれるデジタルプリンタで印字し、封入封緘して郵便物として発送するサービスである。

このサービスにおいて、印刷会社が受取るプリントデータは顧客のコンピュータシステムの文字環境で生成されているので、とくに人名などの外字に関しては文字図形の確認を顧客と印刷会社の間で事前に行い、印刷会社のコンピュータシステムの文字環境を個々の処理作業ごとに顧客のそれに対応させているのが現状である。顧客と印刷会社の双方で標準化された文字図形識別番号が利用できれば、文字図形識別番号を指定するだけで文字図形が確定するので、顧客と印刷会社の間で事前に行う煩雑な文字図形の確認作業が省略できる。

図2 データプリントサービスにおける外字処理



学術出版

東洋・日本に関わる人文学研究（歴史・地理・文学・宗教・思想・科学史など）においては、当該学問が発展してきた歴史的経緯から、古典文献、及び多くの外字・異体字を含む固有名詞（人名・地名・動植物鉱物名など）を扱う機会が多い。しかし、業務システム全般のデジタル化が進

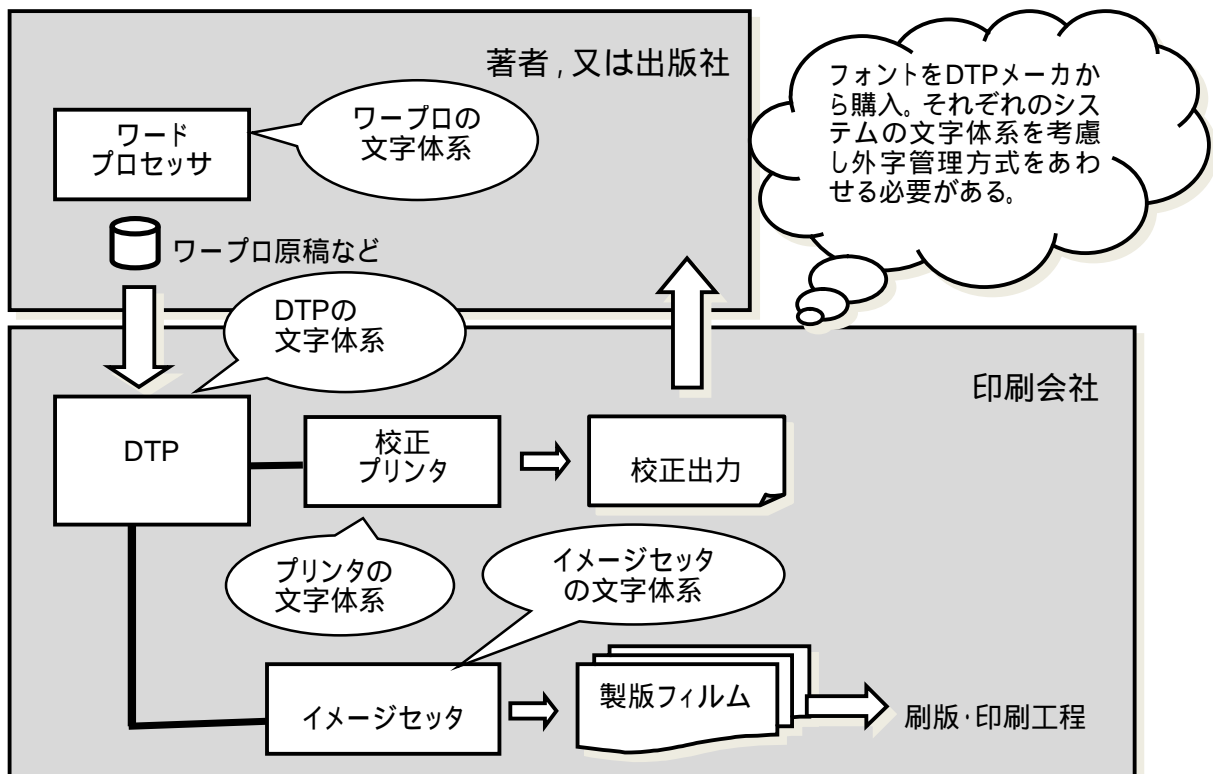
む中であって、それらの外字・異体字を含む出版物のための処理は特例として扱われ、十分な利用環境が整備されているとはいえない。そのため、出版価格の高騰と、成果物公開に際しての大きな阻害要因となっているのが現状である。

この問題は、著者が原稿を書く段階、出版社が原稿を整理・編集する段階、印刷会社が原稿を印刷する段階のそれぞれで利用するワードプロセッサ（以下、ワープロと略す）、デスクトップ・パブリッシング（以下、DTP と略す）、プリンタ、イメージセッタなどにおいて文字情報（文字コード、フォント）を共通に使用できていないことに原因がある。

そのため、データの受け渡し、外字・異体字の処理（ゲタ文字）文字化けに大きな労力を割かざるをえず、作業効率を低下させる要因ともなり、ひいては当該の学問の発展の大きな妨げとなっている。

これらの問題を根本的に解決するには、一連の作業を一貫して流れる文字において、文字図形が固定、共通したものであることが必要である。つまり、文字図形に識別番号を付して形を固定し、標準化することで障害が解決されることになる。

図3 学術出版におけるデータの流れと処理



デジタルアーカイブ

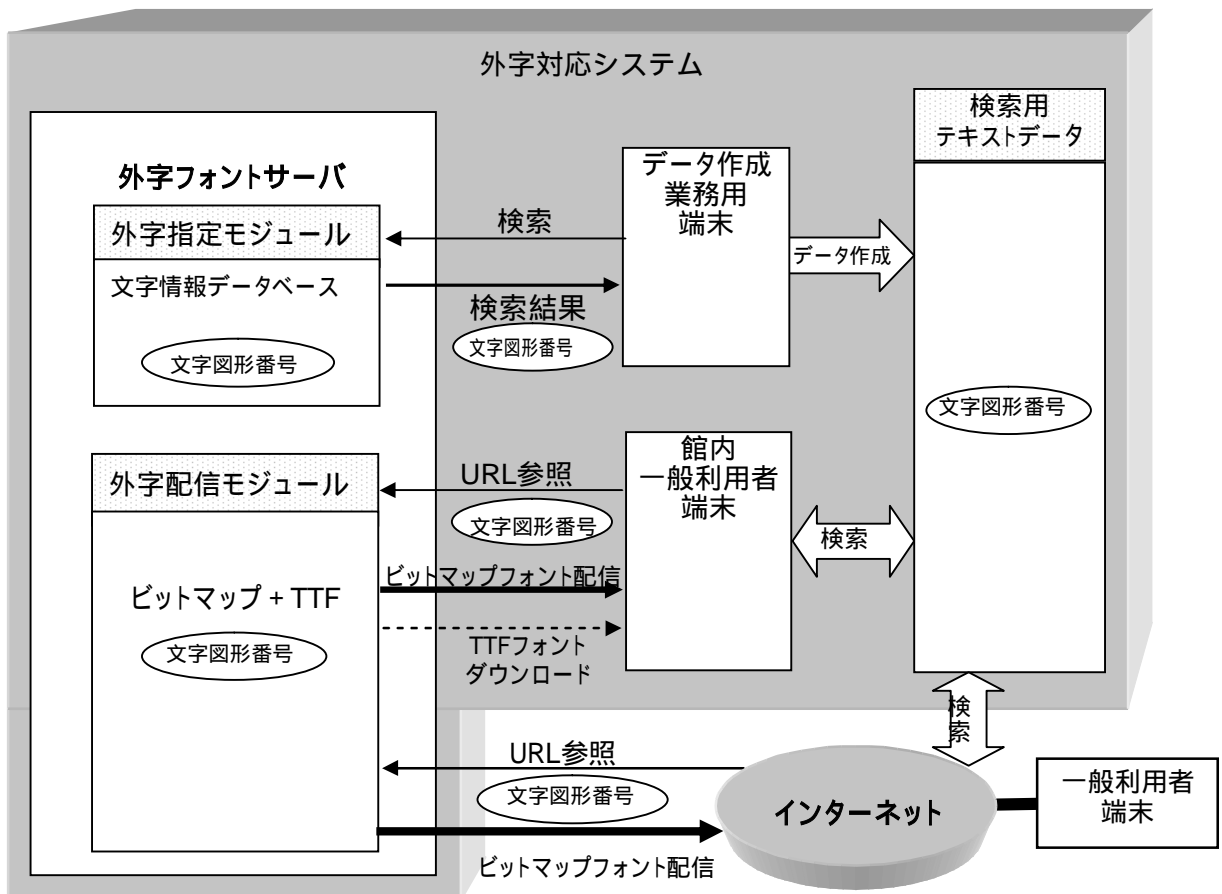
デジタルアーカイブとは、古今東西の知識情報を収集、デジタルデータとして蓄積保存し、広く利活用できるようにしたもので、次に示すようなものがある。

- ・美術館・博物館が保存する文化遺産
- ・大学・研究機関が所蔵する学術資料、及びコレクション
- ・国や自治体の公文書、及び郷土資料

貴重資料もデジタル化することにより、破損・紛失を恐れることなく閲覧提供できるため、日本においても多様な資料がアーカイブ化されてきている。しかしそのデータ構築の手法は、それが構築された時点で用いられていた技術的な制約もあり、単に画像形式化したものから文書の全文電子化まで実に様々であり、特に文書内の外字・異体字問題に対しては共通の解決策がないまま、今日に至っている。

デジタルアーカイブはインターネットを通じた提供が一般的であり、発信者側と利用者側の文字セット環境の違いが文字化け、文字消失の原因となる。この環境の違いを埋めるために、表現すべき文字図形の全てに一意に番号をつけ、その番号を補助的に使うことで、字形の表示や印刷の際に起きる問題の解決が図れ、データ構築の効率化にも大きな効果が見込めるのである。

図4 デジタルアーカイブにおける外字処理



8. 文字図形番号の歴史

本規格案で採用している文字図形集合と文字図形識別便号は、文字鏡研究会がその活動の中心

となる文字図形集合開発プロジェクトが、長年に亘って収録してきた今昔文字鏡の文字図形集合が基本となっている。これはこの文字図形集合が、現存する異体字や外字の集合のなかで最大規模であり、一定の普及をみていることによる。

文字図形集合開発プロジェクトによる文字図形の収録に際しては、文字鏡研究会が組織的に支援してきた。広く一般から受け付けた文字図形作製依頼は、文字鏡研究会の高い水準の学術的な検証を経た上で、文字図形作製を株式会社エーアイ・ネットが行うという形態により、文字図形集合の学術的信頼性が維持されている。

文字鏡研究会では、文字作製依頼受理を自由平等に無償で行うこと、文字作製も同様に無償であること、当該文字図形の非営利使用に対しては対価を要求しないことを条件としてきたため、非営利・学術目的では文字図形集合と文字鏡番号の自由利用が保証されてきた。

文字図形依頼者及び配布を受けた利用者は、立法・司法・行政の三権に所属する複数の組織があり、民間では出版社・印刷会社、銀行・証券・保険会社、テレビ局、及び人名を扱う必要のある法人、研究分野では大学及びそれ以外の研究機関・図書館・博物館・資料館など、国外ではベトナム・台湾・米国・ロシアの研究機関、及びフランス・ドイツの研究者などがいる。

なお、文字鏡文字図形集合は、2006年にISOの図形登録規格であるISO/IEC10036にグリフ識別子として登録された。

9. あとがき

本稿では、JIS化を目指して現在進行中の規格原案「識別番号を持った印刷参照用文字図形の集合」について、その必要性の背景や目的とするところを中心に記述した。

一般に文字に関してはその文化的な背景から様々な領域の方が意見を持ち、規格として定めることになじまない性質のものである。国としても関係省庁は複数にまたがる。しかし、あらゆるデータがコンピュータで処理されることが必要不可欠な現代では、文字の種類を文字コードで表す規格だけではカバーすることが困難な領域がある。本規格原案はこのような限定された領域での利用を想定したものである。

しかし、文字に関する新たな概念の規格を制定することには多くの困難が予想される。何よりも本規格原案の内容を誤解なく理解していただくことが前提となる。あえてJIS化前の原案作成の段階でご説明したのは、関係する方々にご理解頂き、JIS化に向けて是非ご協力いただきたいからである。

なお、本規格原案の作成は、文字図形番号標準化検討委員会およびWGの多くの方のご努力によるものである。小職がその一員として作業に参加し、そしてこの原稿を作成できたことを、本規格原案の作成に関与している全ての方々に感謝する。