文字コードの変遷

一般社団法人 文字情報技術促進協議会 理事・事務局長
一般社団法人情報処理学会 情報規格調査会 規格役員 （ISO/IEC JTC1 NB）
ISO/IEC JTC1 SC2 （Coded character sets） 委員

田丸 健三郎
E-mail: k.tamaru@moji.or.jp

シフトJIS、マイクロソフト標準キャラクタセット、JIS X 0213からUnicode サロゲートペアまで

文字コードの変遷

Windows の変遷

1993

Microsoft
Windows NT Server 3.1

1994

Microsoft
Windows NT Server 3.5

1995

Microsoft
Windows NT Server 3.51

1996

Microsoft
Windows NT Server 4.0

Webサーバーやアプリケーションサーバーとしての機能を強化。
IISやDNS,ASP,DCOMを初めて標準装備、また初めてクラスターに対応

2000

Windows 2000

インターネット標準に対応。
Active DirectoryやPKIの対応など、スケーラビリティとセキュティが飛躍的に向上

2003

Windows Server 2003

初めての64Bit版をリリース。
Webサービスへの対応、Active DirectoryやIISなどを強化

2005

Windows Server 2003 R2

強固なセキュリティで堅牢性を強化。
遠隔拠点のサーバーやID管理を容易にする機能を提供

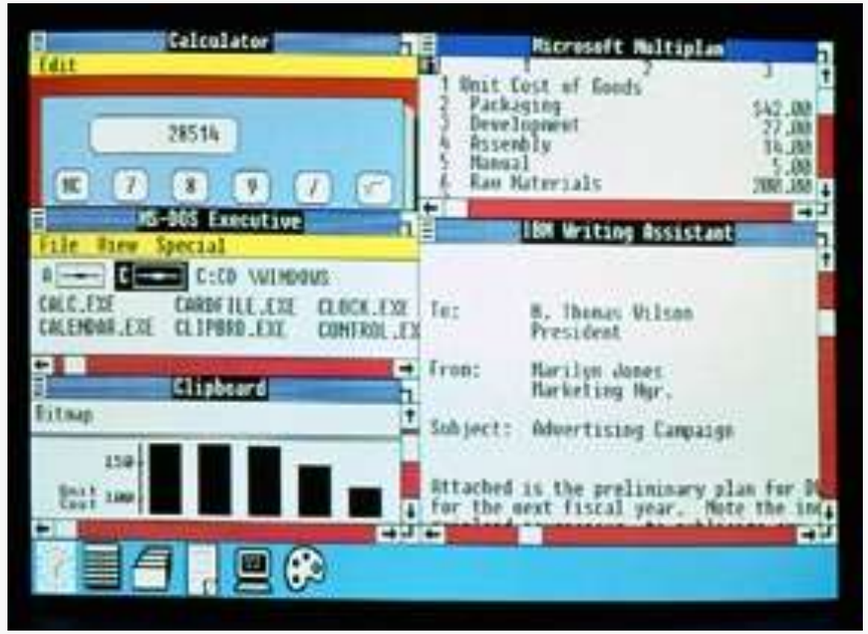
2008

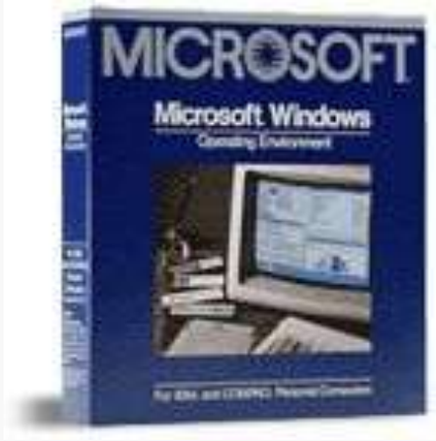
Windows Server 2008

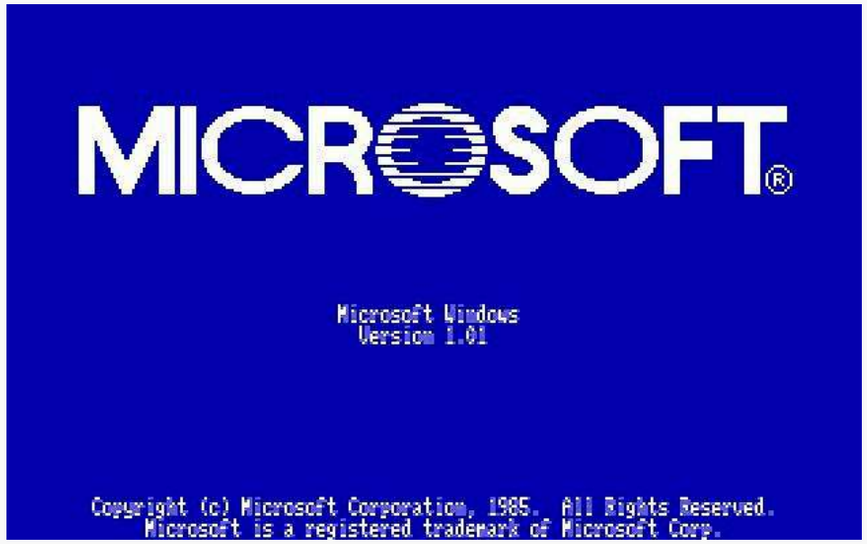
サーバー管理の刷新、仮想化テクノロジの統合、ネットワーク保護、
ターミナルサービスの強化などかつてない管理性、堅牢性、柔軟性を提供

1985年 - Windows 1.0

- 16bit OS
- 初のGUI
- アプリケーションの同時実行

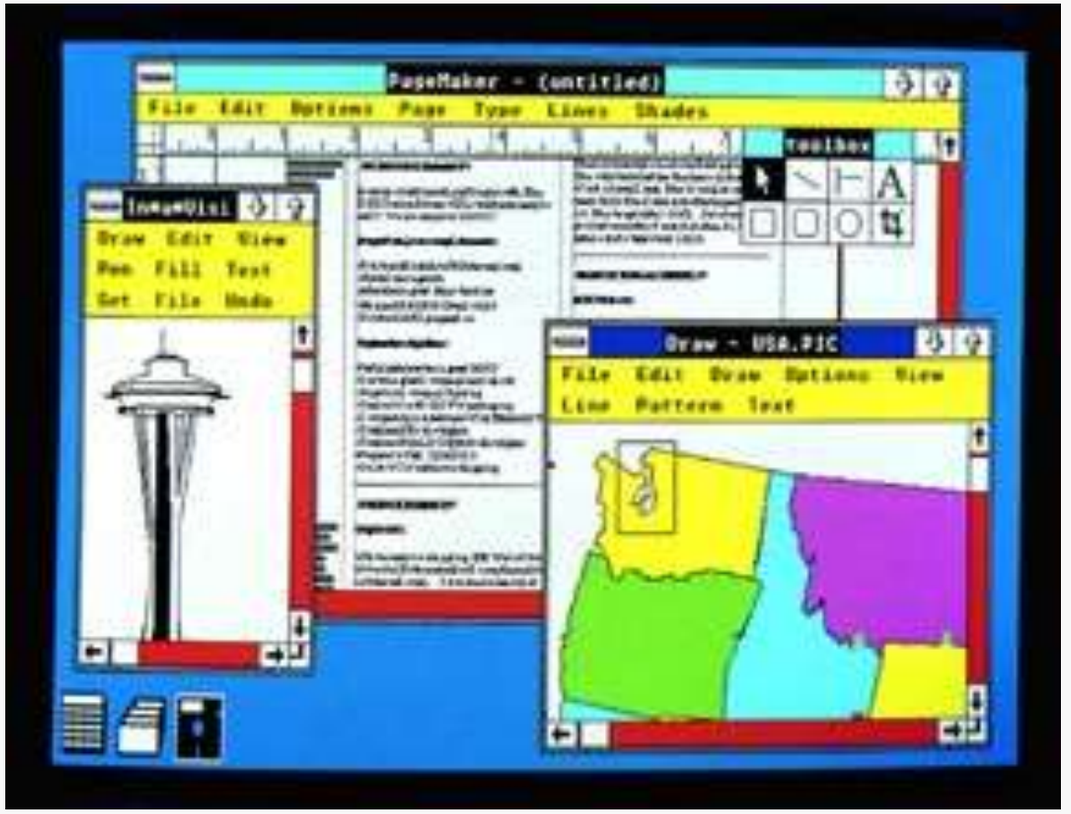






1987年 - Windows 2.0

- キーボードショートカット
- i286プロセッサの採用し物理メモリ空間が
  16MBに拡張。1GBの仮想記憶も使用可能に。
  （多くのPCは1MB未満)
- DDE (Dynamic Data Exchange)
- ウィンドウの重なり

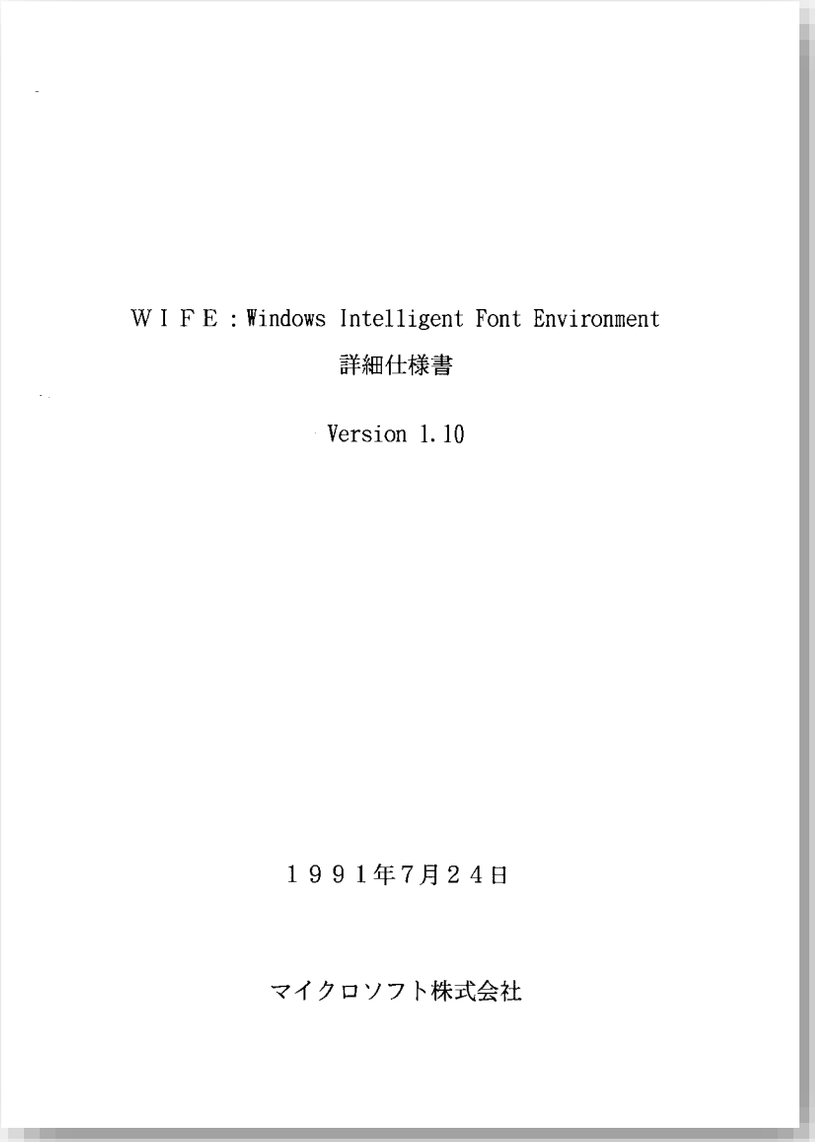


CPUは、フラグレジスタのbit15-12で判定
- 1111bから変更可：8086世代
- 0000bから変更不可：80286
- それ以外：80386以降

WIFE（Windows Intelligent Font Environment）

DBCSフォントを扱うえるようにするために開発された『**インストール可能なフォント・ドライバを実現するサブシステム**』

1990 2月5日　DDS詳細仕様書 V0.1
1990 6月7日　WIFE詳細仕様書 V0.2
1990 6月27日　WIFE詳細仕様書 V0.21
外字等の仕様を追加
1990 7月2日　WIFE詳細仕様書 V0.31
文字の回転（縦書）の仕様を追加



ＷＩＦＥ：Windows Intelligent Font Environment

詳細仕様書

Version 1.10

１９９１年７月２４日

マイクロソフト株式会社

１　概要

１－１　ドキュメントの定義

本ドキュメントは、ＭＳ－ＷＩＮＤＯＷＳ３．Ｘにおいて、インストール可能なフォント・ドライバーを実現するサブシステムの仕様書です。

これは、互いに関連する一群のモジュールによって構成される一つのフォント・サポートの環境であり、この環境の総称を、ＷＩＦＥ（Ｗｉｎｄｏｗｓ　Ｉｎｔｅｌｌｉｇｅｎｔ　Ｆｏｎｔ　Ｅｎｖｉｒｏｎｍｅｎｔ）と呼びます。

ＷＩＦＥ環境では、ＳＢＣＳ文字とＤＢＣＳ文字の区別なくフォントを利用でき、いままで統一されたインストール可能なフォント・システムを持たなかったＤＢＣＳ圏でインストール可能なシステムが利用できるようにするものです。

ＷＩＦＥは、欧米のＳＢＣＳ圏のピュアなＭＳ－ＷＩＮＤＯＷＳ環境に、いくつかのモジュールを付加することによって実現されます。このドキュメントでは、これらの付加されるモジュールと、それによって実現されるＷＩＦＥ独自の環境のみに関して記述しています。そのため、このドキュメントを読むためには、ＭＳ－ＷＩＮＤＯＷＳのＧＤＩモジュールとデバイス・ドライバーとのインターフェース、及び、デバイス・ドライバーの文字列出力機能の詳細に関する知識を必要とします。

１－２　開発の目標

ＷＩＦＥ環境の目標とするのは、次の二つです。

１）ＤＢＣＳフォントを扱うことができる

２）オペレーティング・システムのレベルでサポートする

３）ファイル及びそれ以外の任意の物理メディアによるフォントの提供を可能とする

４）任意の論理フォーマットによるフォントの提供を可能とする

これらの目標を達成するために、インストール可能なフォント・ドライバーという概念を導入します。フォント・ドライバーとは、フォントデータを読み込み、システムが理解できる標準形式のビットマップに変換するプログラムです。このプログラムをユーザーが自由にインストールできるようにすることにより、任意のメディアかつフォーマットによるフォントを利用できるようにします。フォント・ドライバーはフォントデータの提供者が作成し、フォントデータとともに配布することを基本とします。

このフォント・ドライバーを使用可能とするためのサポートモジュール群を含めたシステム全体をＷＩＦＥ環境と呼びます。

このＷＩＦＥ環境を達成するにあたって、次の副目標を設定します。

ａ）ＷＩＮＤＯＷＳアプリケーションは、ＷＩＦＥ環境をサポートするために一切の変更を必要としない。（同じアプリケーションをＷＩＦＥ環境の有無にかかわらず使用できること）

ｂ）ＷＩＮＤＯＷＳカーネル・モジュール群（ＫＥＲＮＥＬ、ＵＳＥＲ、ＧＤＩ）は、ＷＩＦＥ環境の追加によって、一切の変更を必要としない。

ｃ）ＷＩＮＤＯＷＳデバイス・ドライバー（従来から存在するディスプレイ及びプリンターのドライバー）は無変更でＷＩＦＥ環境で使用できること。その際には、パフォーマンスの低下は容認される。

ｄ）前項と関連して、ＷＩＮＤＯＷＳデバイス・ドライバー（ディスプレイ及びプリンターのドライバー）は、ＷＩＦＥ環境に適応するように修正を加えることによって、パフォーマンスを改善できる。

ｅ）ＷＩＮＤＯＷＳシステムの持つ既存の全てのフォントを使用できること。

ｆ）将来的にＷＩＮＤＯＷＳシステムに組み込まれる可能性のあるフォント・システムと可能な限り共存できるよう配慮する。

ｇ）新規に追加されるモジュールであるフォント・ドライバーは、ＯＳ／２のフォントシステムと可能な限り互換性を持たせる。（ソースレベルの互換性）

ｈ）コントロールパネルは、ＷＩＦＥ独特の機能の追加に対応して変更を加える。

ｉ）特定の言語（日本語等）に依存しない。

ｊ）一つのフォント・データで、一つのキャラクターセットの文字を提供する。一つのフォントを、ＳＢＣＳとＤＢＣＳで別々のデータとして提供するようなことはしない。

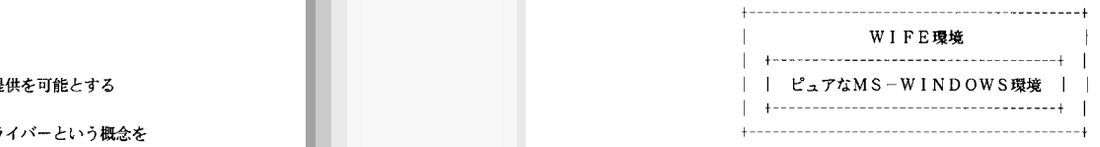
以上。

１－３　互換性

本節では、欧米のＳＢＣＳ圏のピュアなＭＳ－ＷＩＮＤＯＷＳ環境に対する、ＷＩＦＥ環境を含んだＭＳ－ＷＩＮＤＯＷＳシステムの互換性について述べます。

この二つの環境の関係は、以下の図で示すように、ピュアなＭＳ－ＷＩＮＤＯＷＳの全てを含み、それを外側から包むような形で、ＷＩＦＥ環境が存在します。そのため、ＷＩＦＥ環境では、ピュアなＭＳ－ＷＩＮＤＯＷＳ環境の機能は全て使用できます。

ですから、ＷＩＦＥ環境は、ピュアなＷＩＮＤＯＷＳ環境の上位環境であると位置づけられます。



なお、上の図は、フォントの取り扱いの概念について説明したものです。

実際には、ＵＳバージョンなどのピュアなＭＳ－ＷＩＮＤＯＷＳに、モジュールを追加することで、ＷＩＦＥ環境を実現することは不可能です。なぜなら、ピュアなＭＳ－ＷＩＮＤＯＷＳ環境を構成するモジュールには、ＤＢＣＳキャラクターを正しく扱うための能力がないためです。ＤＢＣＳフォントをインストール可能にすることと、ＤＢＣＳキャラクターを正しく扱うことは、全く別の問題です。本仕様書は、前者の「ＤＢＣＳフォントをインストール可能にすること」についてのみ、対象とします。

１－３－１　ＡＰＩの互換性

アプリケーション・プログラムから呼び出すことができるＡＰＩには、一部の例外を除き、ピュアなＭＳ－ＷＩＮＤＯＷＳ環境に対して、追加や変更はありません。

つまり、既存のＡＰＩによって、ＷＩＦＥ環境の全ての機能が利用可能です。

アプリケーションから見ると、ＷＩＦＥフォントは、デバイス・ドライバーに固有のデバイス・フォントと同等に見えます。そのため、アプリケーションは、ピュアなＭＳ－ＷＩＮＤＯＷＳが持っている全てのデバイス・フォントを利用するためのＡＰＩを利用し、ＷＩＦＥフォントにアクセスすることができます。もちろん、本来のデバイス・フォントも平行して利用できます。

相互運用性の喪失

各ベンダーによりJISが独自拡張されたことにより、異なるメーカー間での文字情報の交換・共有が困難になる。

「マイクロソフト標準キャラクタセット」

各ベンダーによる独自拡張された文字を整理統合し、異機種・プログラム間の文字情報の相互運用を可能に。

Microsoft Windows™ version 3.1
日本語版

マイクロソフト標準キャラクタセット
仕様書

Revision 1.0
Dec 25, 1992

Microsoft Co., Ltd.
Systems R&D

図 A-1 DBCS コードレイアウト

ShiftJIS	JIS (区)	
8140	1	JIS 非漢字 X0208 JIS90
:	:	
8440	7	
:	8	
:	:	未使用 (376 文字)
8740	13	NEC 特殊文字 (83 文字)
:	:	未使用 (188 文字)
8840	15	
:	16	JIS 漢字 X0208 JIS90
:	:	
:	:	
:	:	
:	:	
EA40	83	
:	84	
EB40	85	未使用 (376 文字)
:	:	
ED40	89	NEC IBM 選定文字 (374 文字)
:	90	
EE40	91	
:	92	
EF40	93	未使用 (188文字)
:	94	
F040	95	外字 (1880 文字)
:	:	
:	:	
F940	113	
:	114	
FA40	115	IBM 拡張文字 (388 文字)
:	:	
FC40	119	

US-ASCIIの穴を縫って作られたシフトJIS

文字を指すポインタが文字の先頭か中間か容易に判別できない。

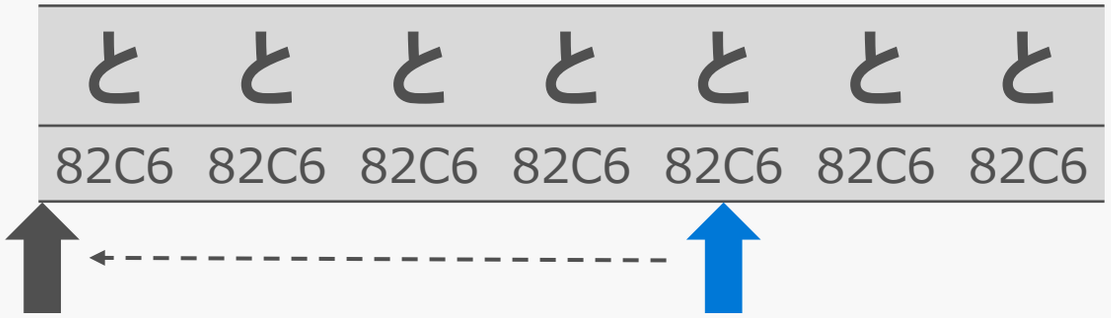
English	と	日	本	語
45 6E 67 9C 69 73 68	82C6	93FA	967B	8CEA

0xC6は文字の先頭か中央か？

JIS X 0201 / JIS X 0208 / JIS X 0213	第 1 バイト	第 2 バイト
1 バイト文字　ASCII (SBCS / Single Byte Character Set)	0x21 ~ 0x7E	
1 バイト文字　半角カタカナ (SBCS / Single Byte Character Set)	0xA1 ~ 0xDF	
2 バイト文字　JIS X 0208 / JIS X 0213（1 面） (DBCS / Double Byte Character Set)	0x81 ~ 0x9F 0xE0 ~ 0xEF	0x40 ~ 0x7E 0x80 ~ 0xFC
2 バイト文字　ユーザー定義文字 (DBCS / Double Byte Character Set)	0xF0 ~ 0xFC	0x40 ~ 0x7E 0x80 ~ 0xFC
2 バイト文字　JIS X 0213（2 面） (DBCS / Double Byte Character Set) JIS　X 0213 にて規定されていますが、Windows をはじめとする多くのプラットフォーム、ツールでは JIS X 0213 の文字としては認識せず、外字として認識します。	0xF0 ~ 0xFC	0x40 ~ 0x7E 0x80 ~ 0xFC

実装コストが高く、しばしばバグの原因に

- 文字列のポインタをそのまま使用するバグる
- 1文字前のポインタを取得するだけでも計算資源を浪費する



文字列の先頭まで辿らないと文字の先頭か「ニ」（0xC6）か判別できない

```
LPSTR SzAlign(LPSTR szStart, LPSTR szSource)
{
  LPSTR sz;
  CPINFO  cpinfo;

  if (!szStart || !szSource)
     return NULL;

  if (szStart == szSource)
     return szStart;

  sz = szSource;
  // 以下のコードでは、ポインタが確実にDBCSの第1バイトでない
  // 文字が現れるまでポインタを前方移動します。これは、DBCSの
  // 第1バイトと第2バイトの範囲に重複があるためです。
  for(;;)
  {
     if( sz==szStart || !IsDBCSLeadByte(*(sz-1)))
        return (LPSTR)( szSource -
           (((ULONG_PTR)sz^(ULONG_PTR)szSource)&1) );
     sz--;
  }

}
```

0x5C 問題

バックスラッシュはデリミタとして広く使用されている

＼ <Folder> ＼<FileName>

同じコードが日本語環境だと

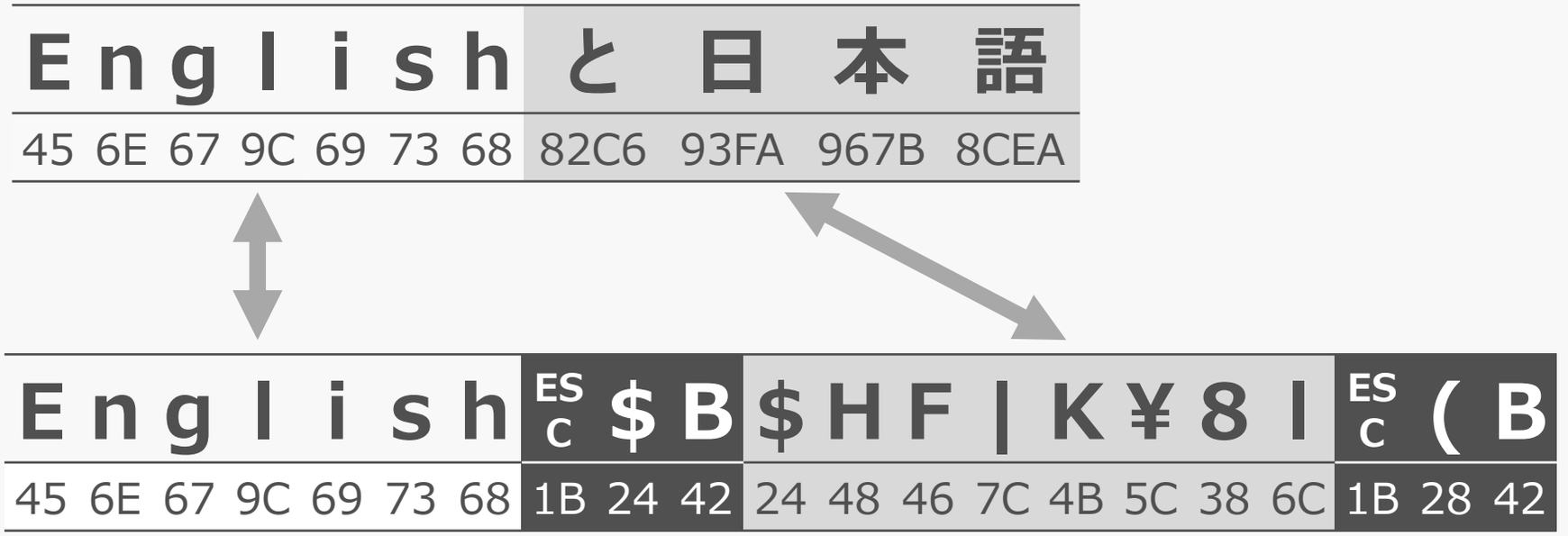
¥

JIS X 0201 / JIS X 0208 / JIS X 0213	第 1 バイト	第 2 バイト
1 バイト文字　ASCII (SBCS / Single Byte Character Set)	0x21 ~ 0x7E	
1 バイト文字　半角カタカナ (SBCS / Single Byte Character Set)	0xA1 ~ 0xDF	
2 バイト文字　JIS X 0208 / JIS X 0213（1 面） (DBCS / Double Byte Character Set)	0x81 ~ 0x9F 0xE0 ~ 0xEF	0x40 ~ 0x7E 0x80 ~ 0xFC
2 バイト文字　ユーザー定義文字 (DBCS / Double Byte Character Set)	0xF0 ~ 0xFC	0x40 ~ 0x7E 0x80 ~ 0xFC
2 バイト文字　JIS X 0213（2 面） (DBCS / Double Byte Character Set) JIS　X 0213 にて規定されていますが、Windows をはじめとする多くのプラットフォーム、ツールでは JIS X 0213 の文字としては認識せず、外字として認識します。	0xF0 ~ 0xFC	0x40 ~ 0x7E 0x80 ~ 0xFC

メールで一般的だった 7bit JIS / ISO-2022-JP

メールの送信ではエンコード、受信ではでコード処理が必要。

当時は方言、送信側バグなどへの対応により高コストの象徴



1978 通産省により定められた最初の文字コード規格

6,802字

1983 規格票例字体 約２５０字に変更

6,877字

1990 第 2水準漢字に 2文字追加

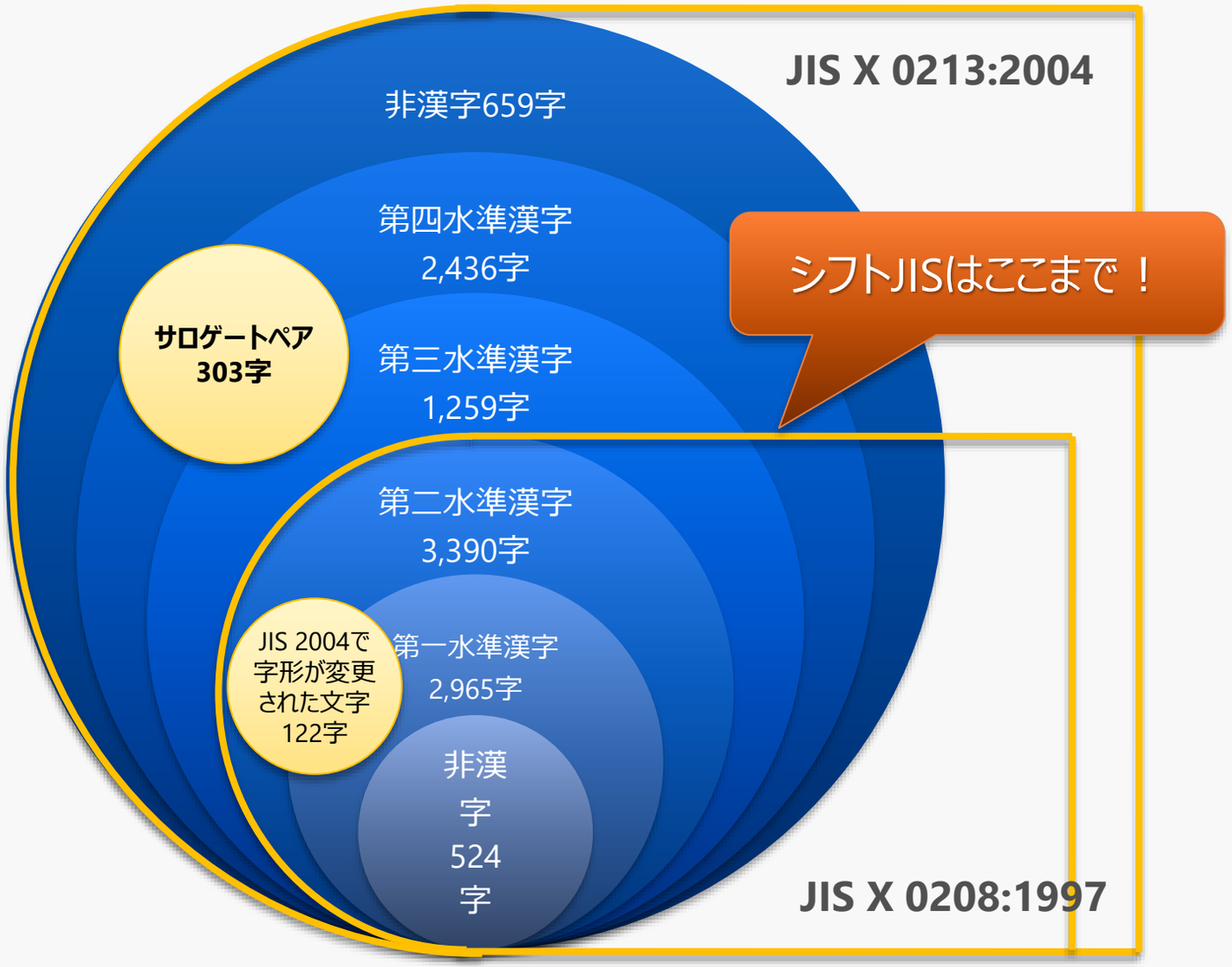
6,879字

2004 印刷標準字体

11,233字

Unicodeの対応

サロゲートペアの使用



Unicodeとサロゲートペア

Windows 95（1995）

- 32bit OS
- Intel Pentium Pro 150～200MHz
- **平均4～16MB メモリ**
- MP (Multi Processor)

二つの議論

1993年夏
Windows 95
でUCS？

第一世代（1983 - ）Shift JIS

(JIS78 or JIS83) + メーカー拡張	MS-DOS 2 - , MS OS/2 1.x, MS-Windows 2.x, 3.0

第二世代(1993 - ）マイクロソフト 標準 キャラクタ セット

JIS90, 10646	Windows 3.1, NT 3.1/3.5, 95, NT 3.51/4.0

第三世代（1998 - ）補助漢字をUCS拡張

補助漢字	Windows 98, NT 4.0 SP4, 2000, Me, XP

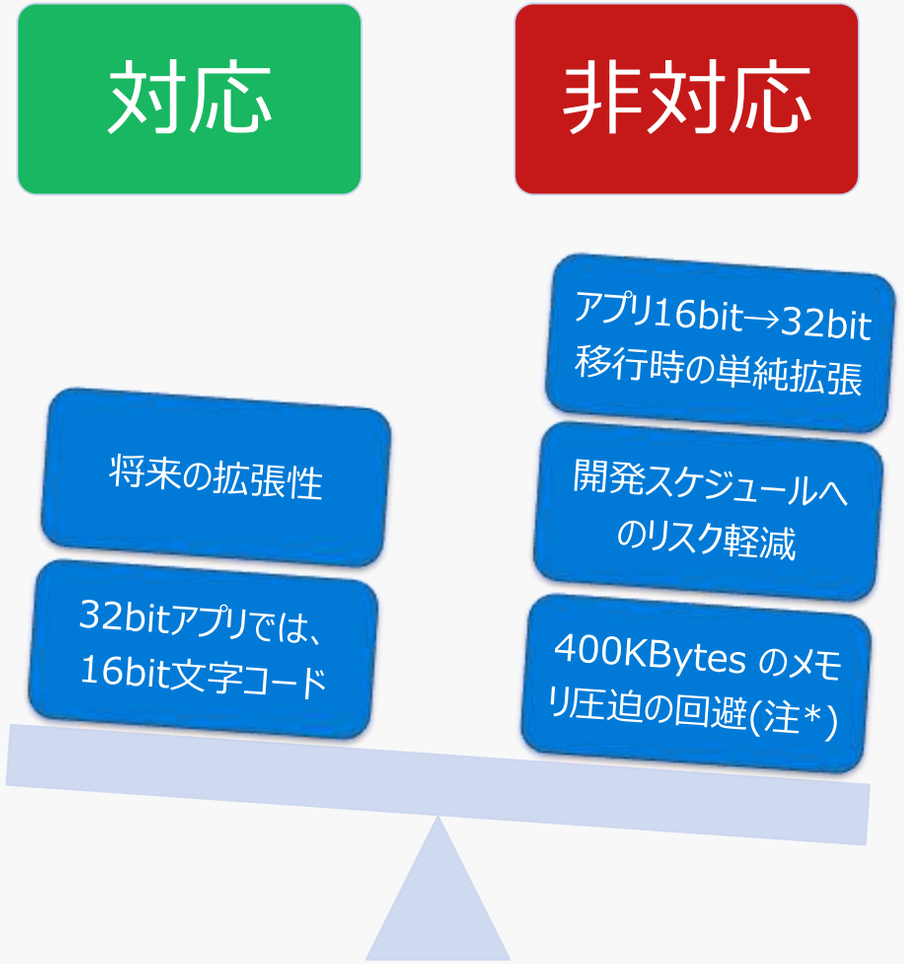
第四世代（2007 - ）国語審議会答申に基づく最新規格への対応

JIS2004	Windows Vista, Windows Server Windows XP (with JIS2004 Pack)

2000年冬
BMP か
Surrogateか？

Windows 95 での UCS 対応

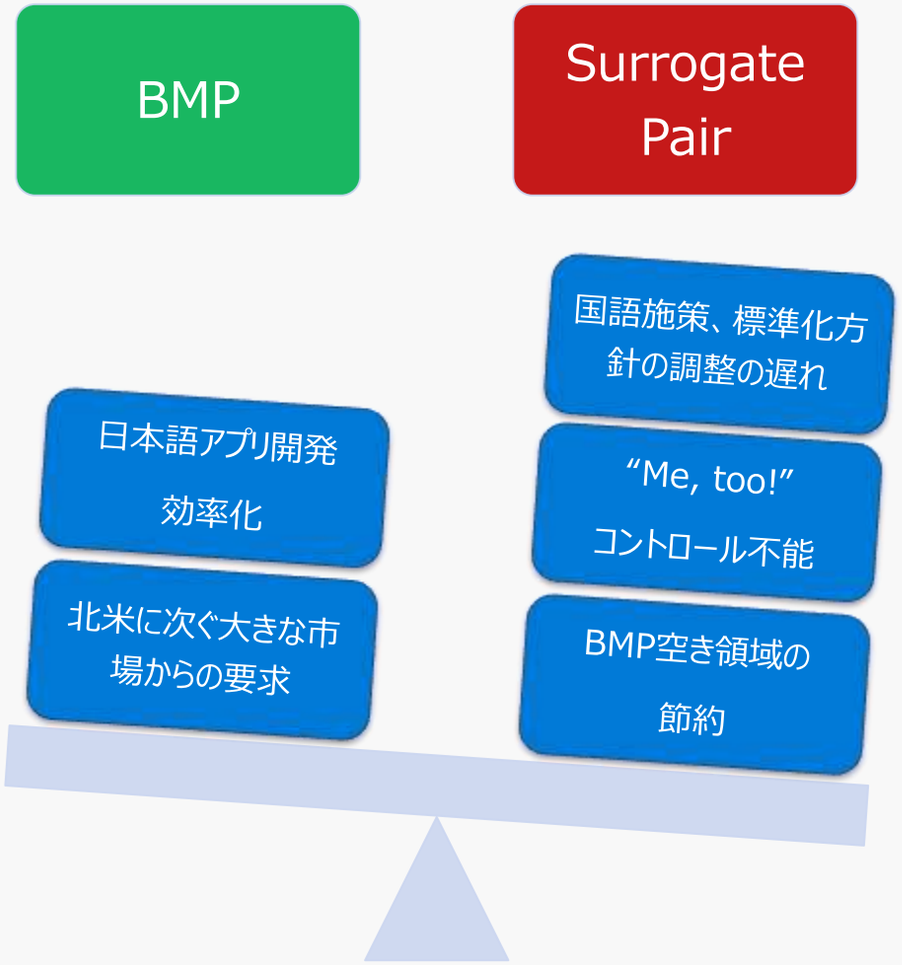
1993年夏 @ 米国 ワシントン州



注*: 2 ~ 4Mbytes の必要メモリを想定

JIS2000 の 303文字の行方

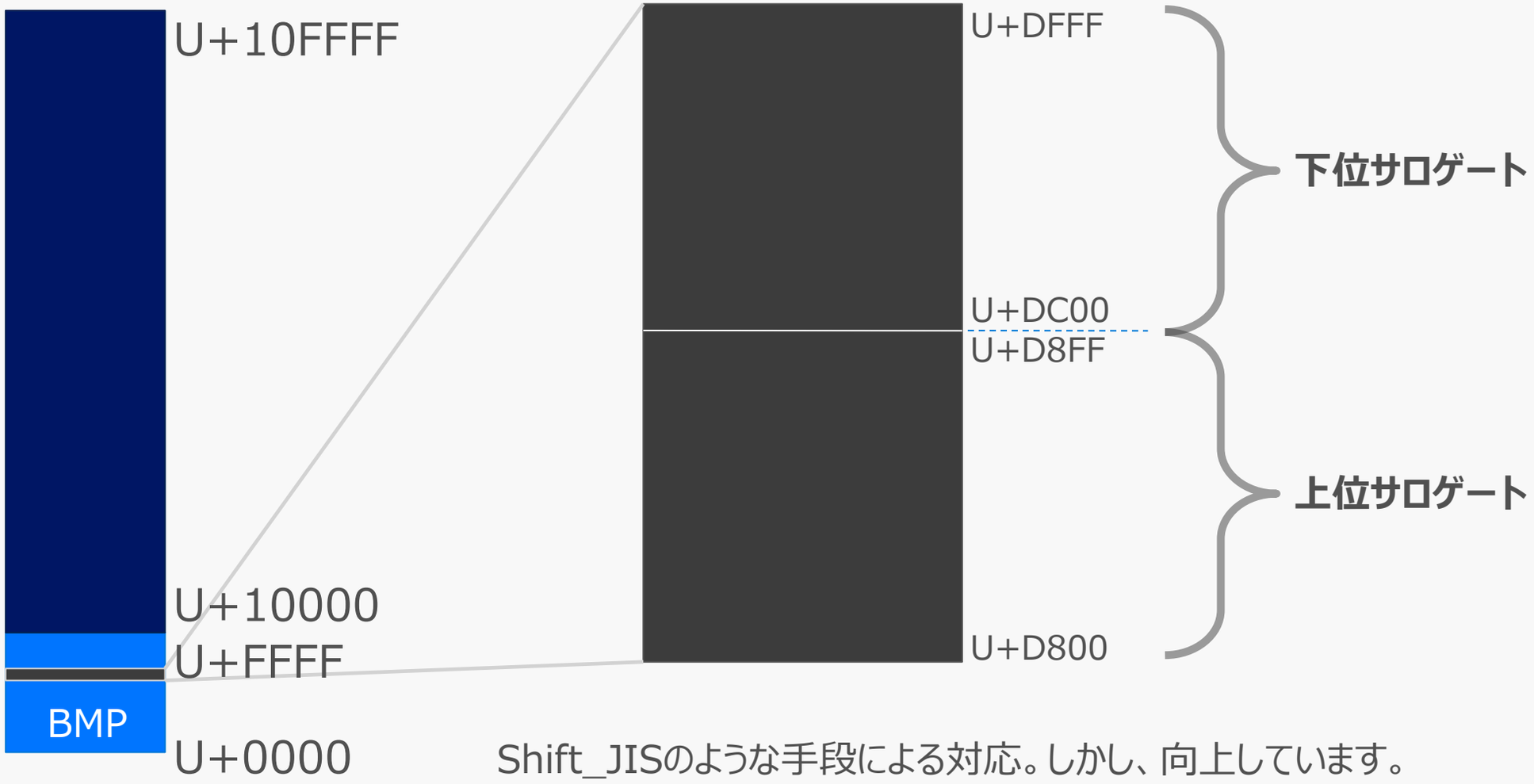
2000年 冬@ 米国 カリフォルニア州、中国 北京



JIS2000により 1 文字 32ビットへの対応が必須に。

（UTF-16 & JIS X 0213:2000）

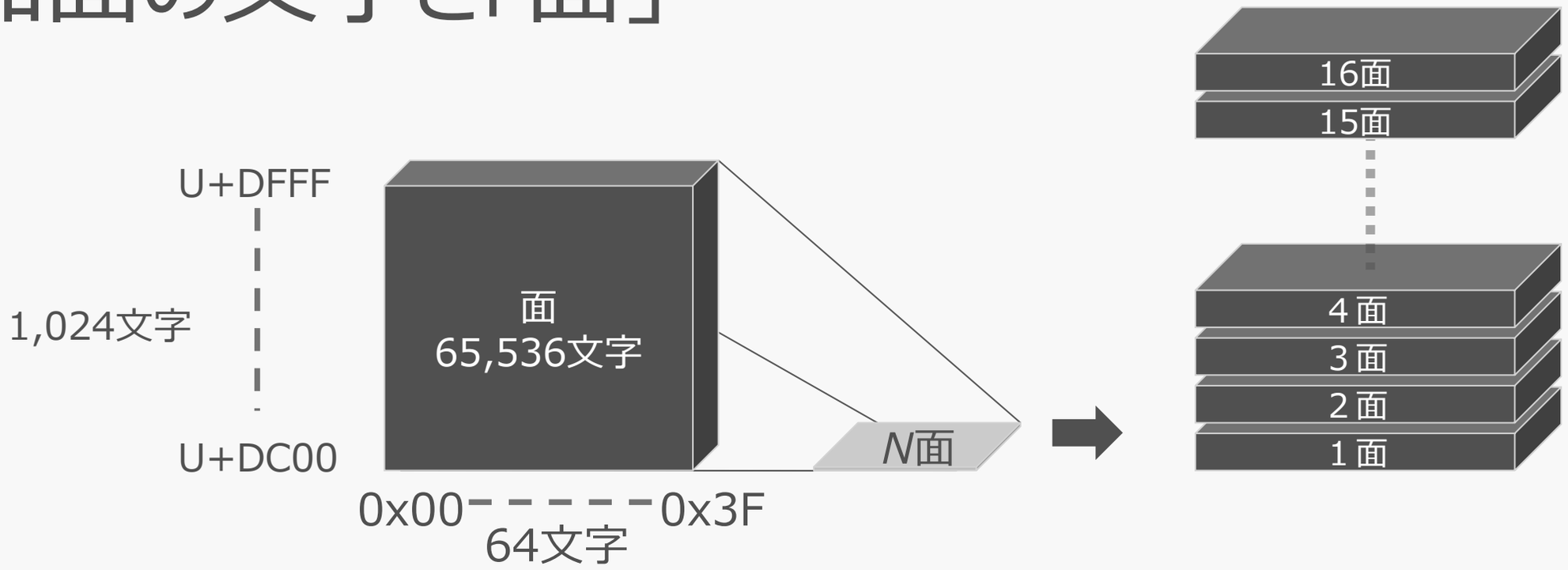
追加面の文字（サロゲートペア）



Shift_JISのような手段による対応。しかし、向上しています。

値から上位サロゲート、下位サロゲートの判断が可能。

追加面の文字と「面」



「	丈	丄	堅	」	は	追	加	面	の	文	字
300C	D840 DC0B	D844 DE3D	D844 DF1B	300D	306F	8FFD	52A0	9762	306E	6587	5B57

上位サロゲート (U+D800 ~ U+DBFF)		下位サロゲート (U+DC00 ~ U+DFFF)		合計拡張文字数
1,024	X	1,024	=	1,048,576文字

JIS2004における大きな変更

- 字形の変更
- 追加面の文字（サロゲートペア）の使用
- Windows Vista, Windows Server 2008はJIS 2004の字体をUIに使用。
- アプリケーションはJIS90, JIS04両方を使用可能(OpenType)

Windows Vista, Windows Server 2008以降に搭載されているOpenTypeフォントは、両方の字形をサポート

書籍では

	使用されている字形数	Unicode		該当無
		Plain	IVD(AJ1-6)	
萬葉集	3,531	99.4%	99.8%	0.2%
梶井基次郎全集	4,563	97.3%	99.8%	0.2%
中島敦全集	3,543	97.6%	99.8%	0.2%

- JISでサポートされている文字（約2,000)だけでは書籍で使用されている字形全てを表現することができない。
- 約97%以上はシフトJIS （CP932）に含まれる
- Unicode、AJ1-6で 約99.8%を網羅

平成22年度書籍等デジタル化推進事業
（デジタル・ネットワーク社会における出版物の利活用推進のための外字・異体字利用環境整備事業）

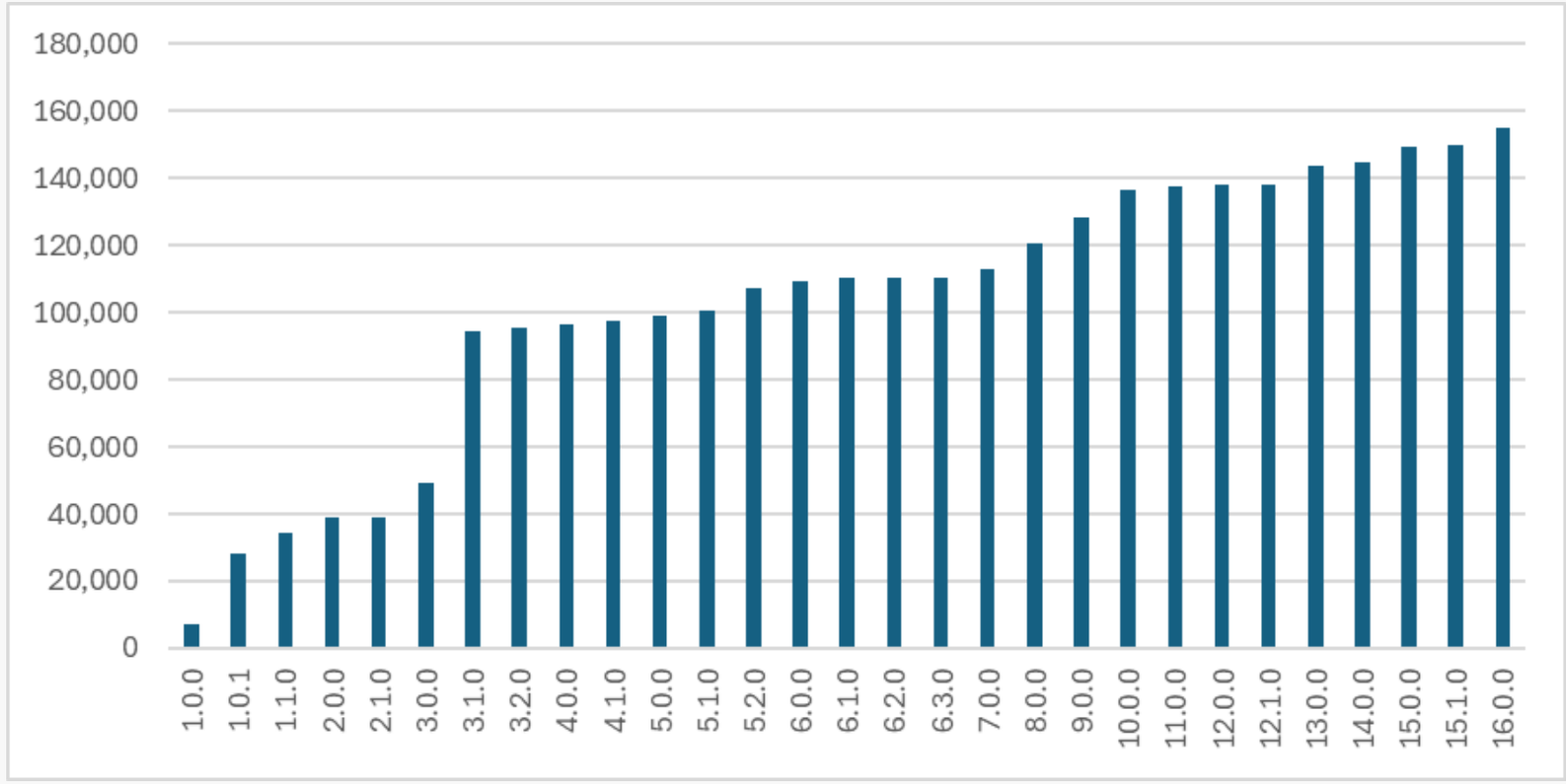
Unicodeの文字数

Unicode バージョン	制定年	文字数	詳細
1.1	1992	34,233	• JIS X 0208 と JIS X 0212 を含む Unicode のバージョン
2.0	1996	38,950	• サロゲート ペアを技術仕様として採用 （この時点では文字は未定義であり、3.1にて実装） • ハングル文字の移動 (Unicode 1.1 と互換性消失) • (技術仕様としては、JIS X 0213:2004 に対応)
2.1	1998	38,952	• ユーロ通貨記号追加、多少数の記号定義変更
3.0	1999	49,259	• CJK 統合漢字拡張 A、漢字 6,582 文字追加
3.1	2001	94,205	• サロゲート ペア 303 文字を追加 • JIS X 0213:2000 一部対応、言語タグ追加 • CJK 統合漢字拡張 B ブロック追加 • CJK 統合漢字拡張 B、漢字 42,711文字追加
3.2	2002	95,221	• JIS X 0213:2000 および JIS X 0213:2004 に正式対応、 • 異体字セレクタ 1 ～ 16 追加 • （JIS X 0213:2004 の追加 10 文字は、すでに存在） • CJK 互換漢字ブロックに追加された JIS X 0213:2000 漢字の 59 文字および追加丸付き数字 (～ ㊿) などの非漢字を追加
4.0.0	2003	96,447	• 異体字セレクタ 17 ～ 256 追加
5.0.0	2006	99,089	• BMP(基本多言語面) 領域にバリ文字など追加 • サロゲート領域にフェニキア文字など追加
6.0	2010	109,449	• ISO/IEC 10646:2010 • 絵文字の追加

それでも足りない文字

Unicodeの登録文字数の変遷

制定年月日	バージョン番号	収録文字数
1991年10月	1.0.0	7,161
1992年6月	1.0.1	28,359
1993年6月	1.1.0	34,233
1996年7月	2.0.0	38,950
1998年5月	2.1.0	38,952
1999年9月	3.0.0	49,259
2001年3月	3.1.0	94,205
2002年3月	3.2.0	95,221
2003年4月	4.0.0	96,447
2005年3月31日	4.1.0	97,720
2006年7月14日	5.0.0	99,089
2008年4月4日	5.1.0	100,713
2009年10月1日	5.2.0	107,361
2010年10月11日	6.0.0	109,449
2012年1月31日	6.1.0	110,181
2012年9月26日	6.2.0	110,182
2013年9月30日	6.3.0	110,187
2014年6月16日	7.0.0	113,021
2015年6月17日	8.0.0	120,737
2016年6月21日	9.0.0	128,172
2017年6月20日	10.0.0	136,690
2018年6月5日	11.0.0	137,374
2019年3月5日	12.0.0	137,928
2019年5月7日	12.1.0	137,929
2020年3月10日	13.0.0	143,859
2021年9月22日	14.0.0	144,697
2022年9月13日	15.0.0	149,186
2023年9月12日	15.1.0	149,813
2024年9月10日	16.0.0	154,998



基本多言語面 (Basic Multilingual Plane)
追加面 (Supplementary Plane)

U+10FFFF

U+EFFFD

Plane 14 (Special-Purpose Plane)

U+E0000

U+3FFFD

Plane 3 (Tertiary Ideographic Plane)

U+30000

U+2FFFD

Plane 2 (Supplementary Ideographic Plane)

U+20000

U+10FFD

Plane 1 (Supplementary Multilingual Plane)

U+10000

U+FFFF

Plane 0 (Basic Multilingual Plane)
基本多言語面

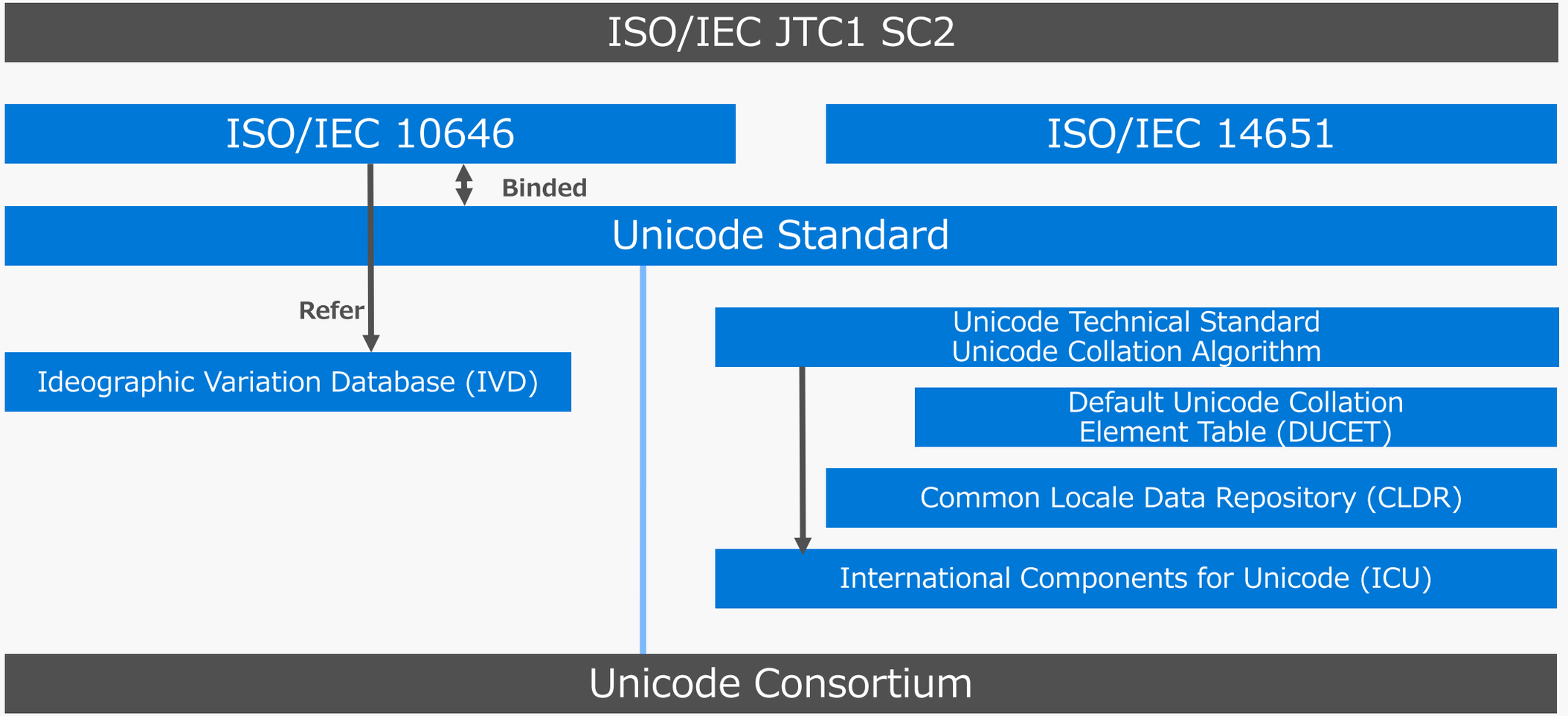
U+0000

Unicodeのアドレス空間は21ビット

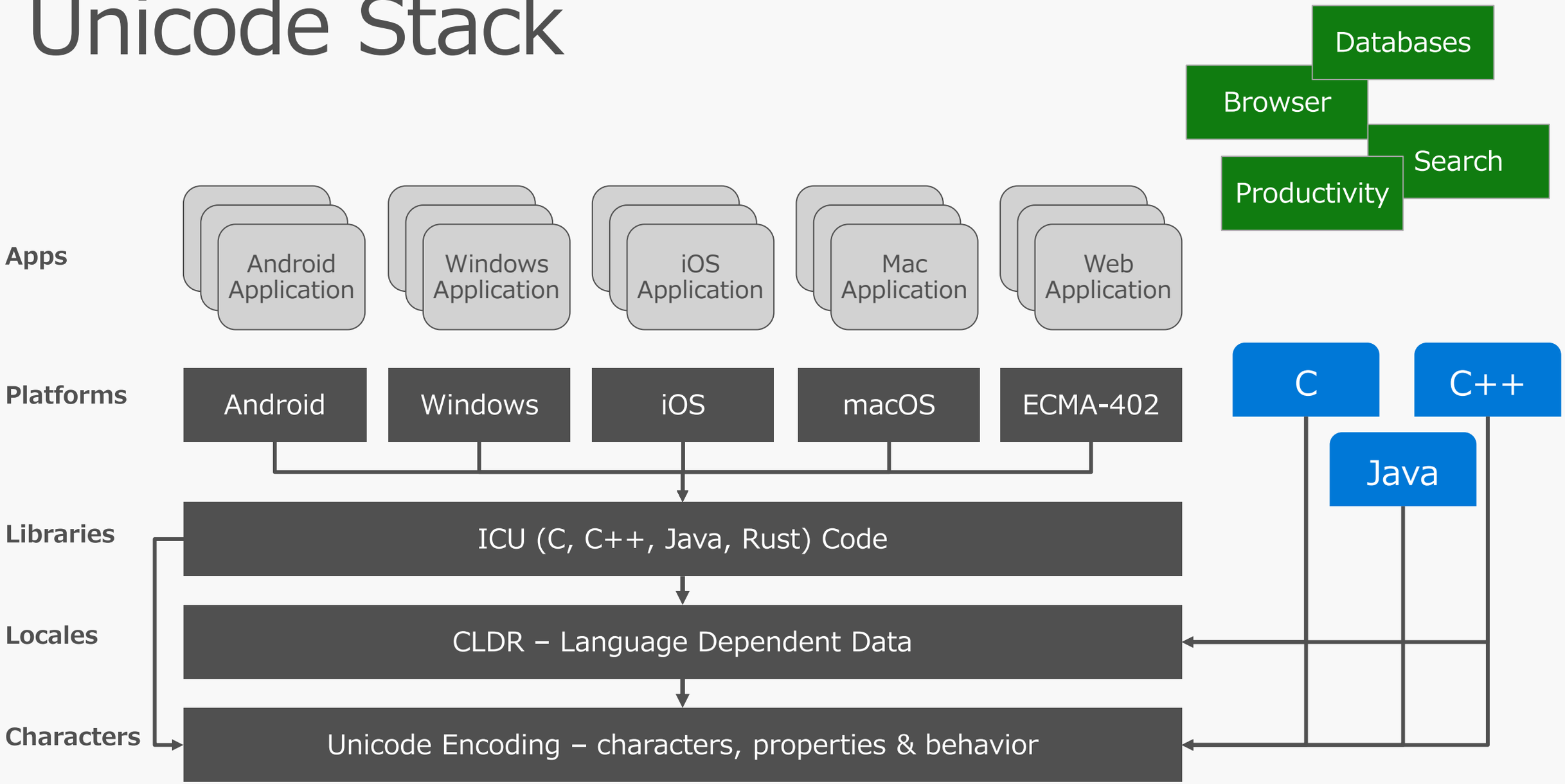
http://unicode.org/roadmaps/index.html

文字と規格の関係

- 国際規格（デジュール）と業界標準（フォーラム）が連携
- 文字に関係する規格は、ICUに集約

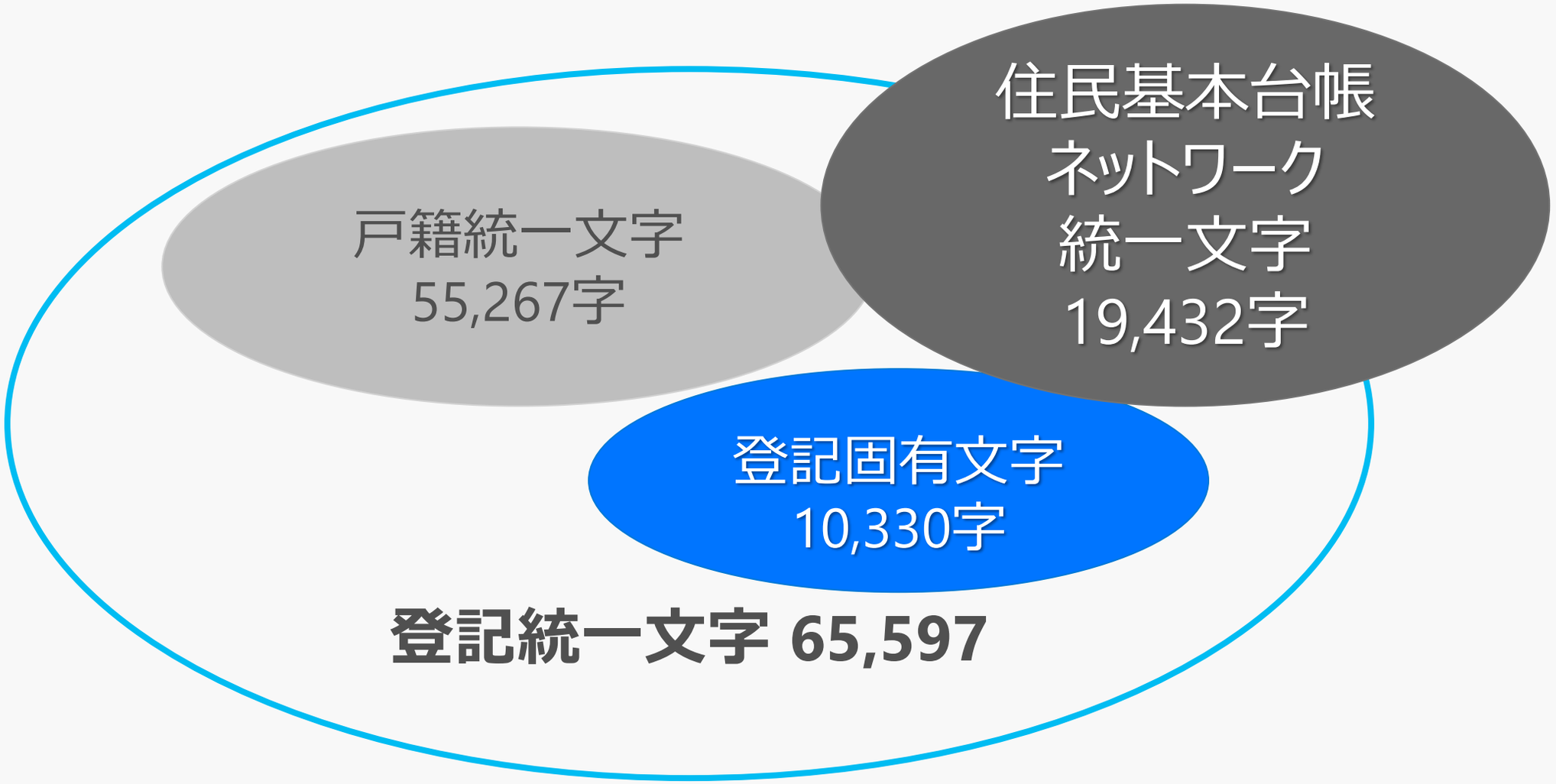


Unicode Stack



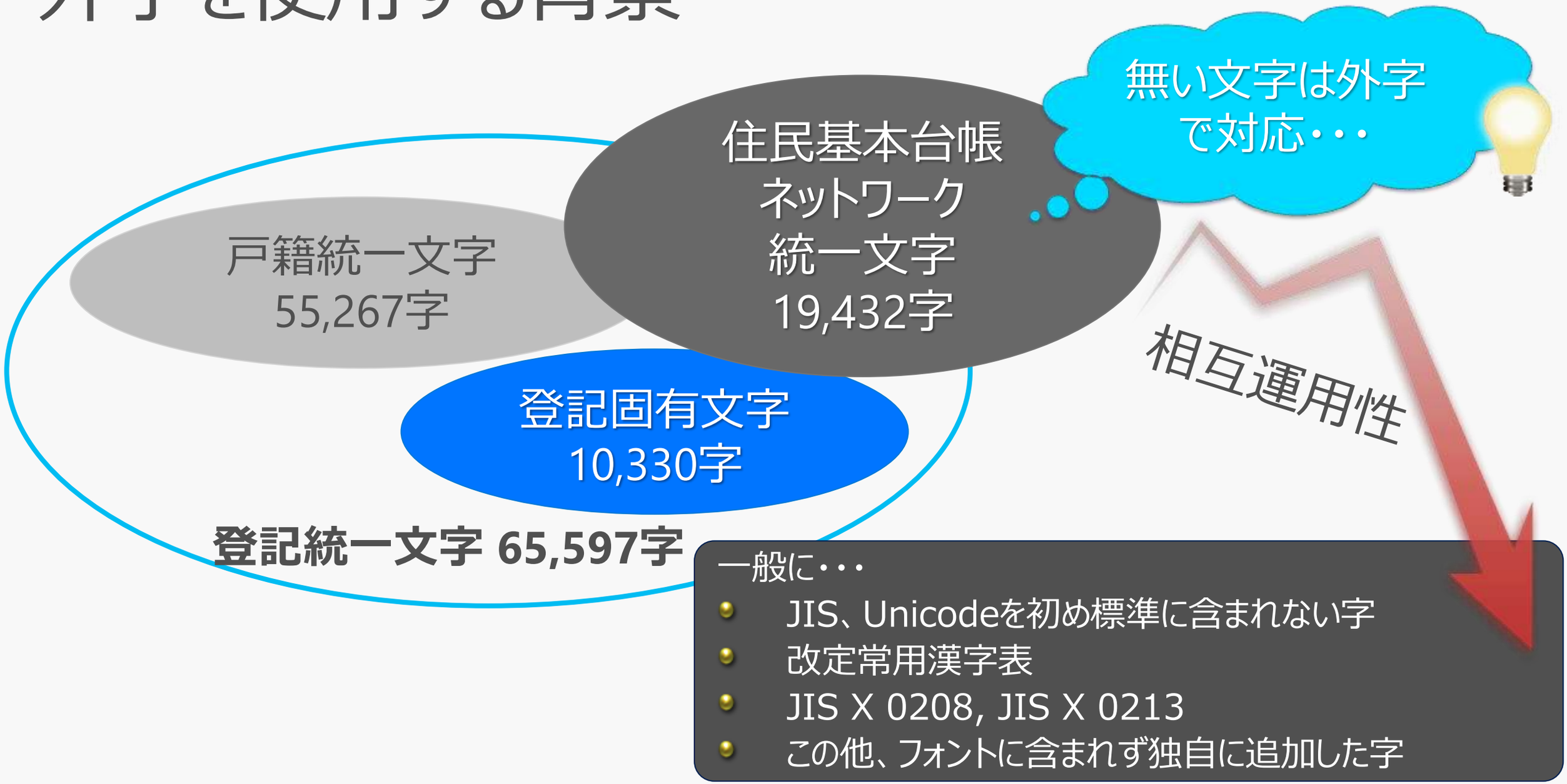
外字問題とIVD

行政システムでは

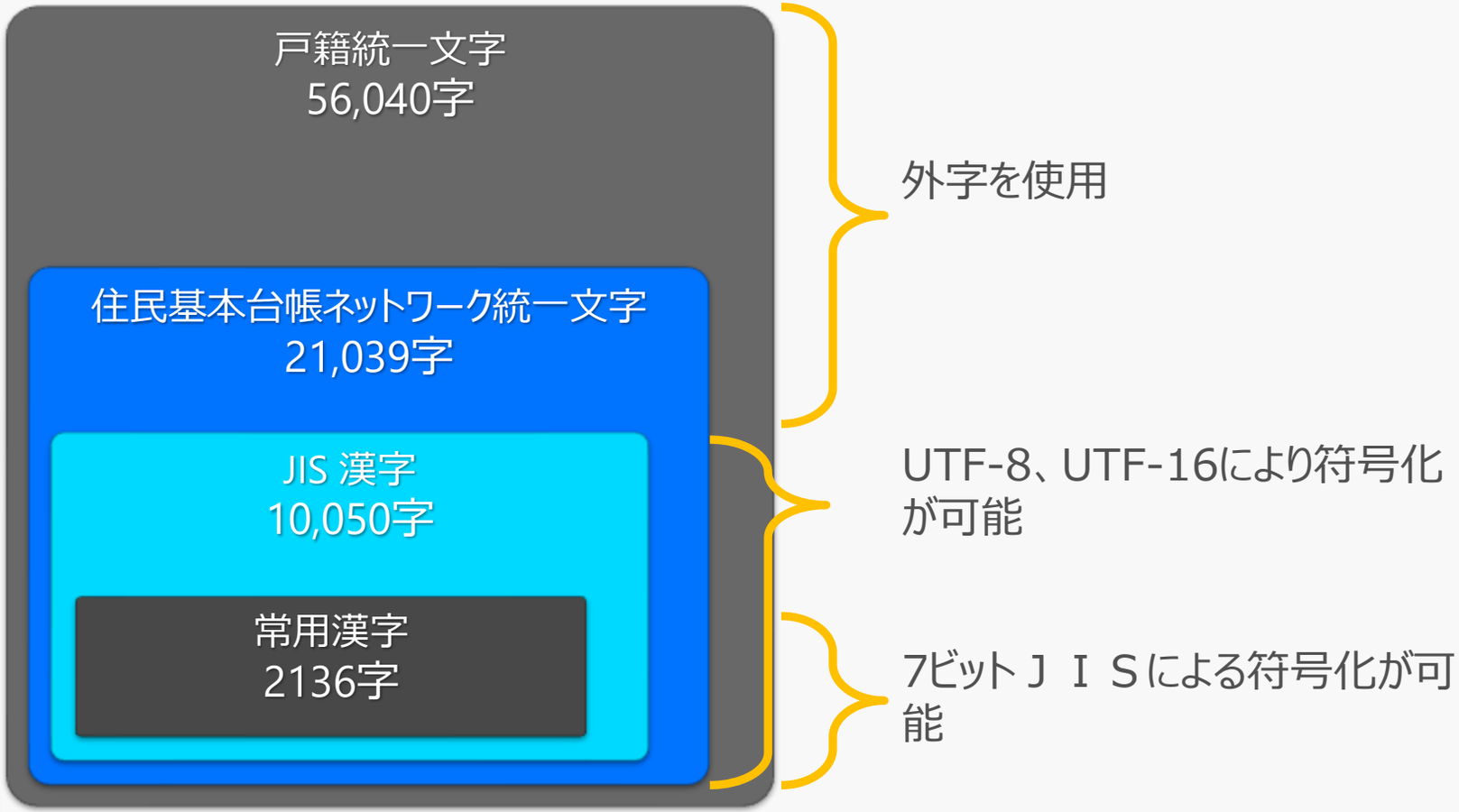


※それぞれ漢字以外の文字を除く

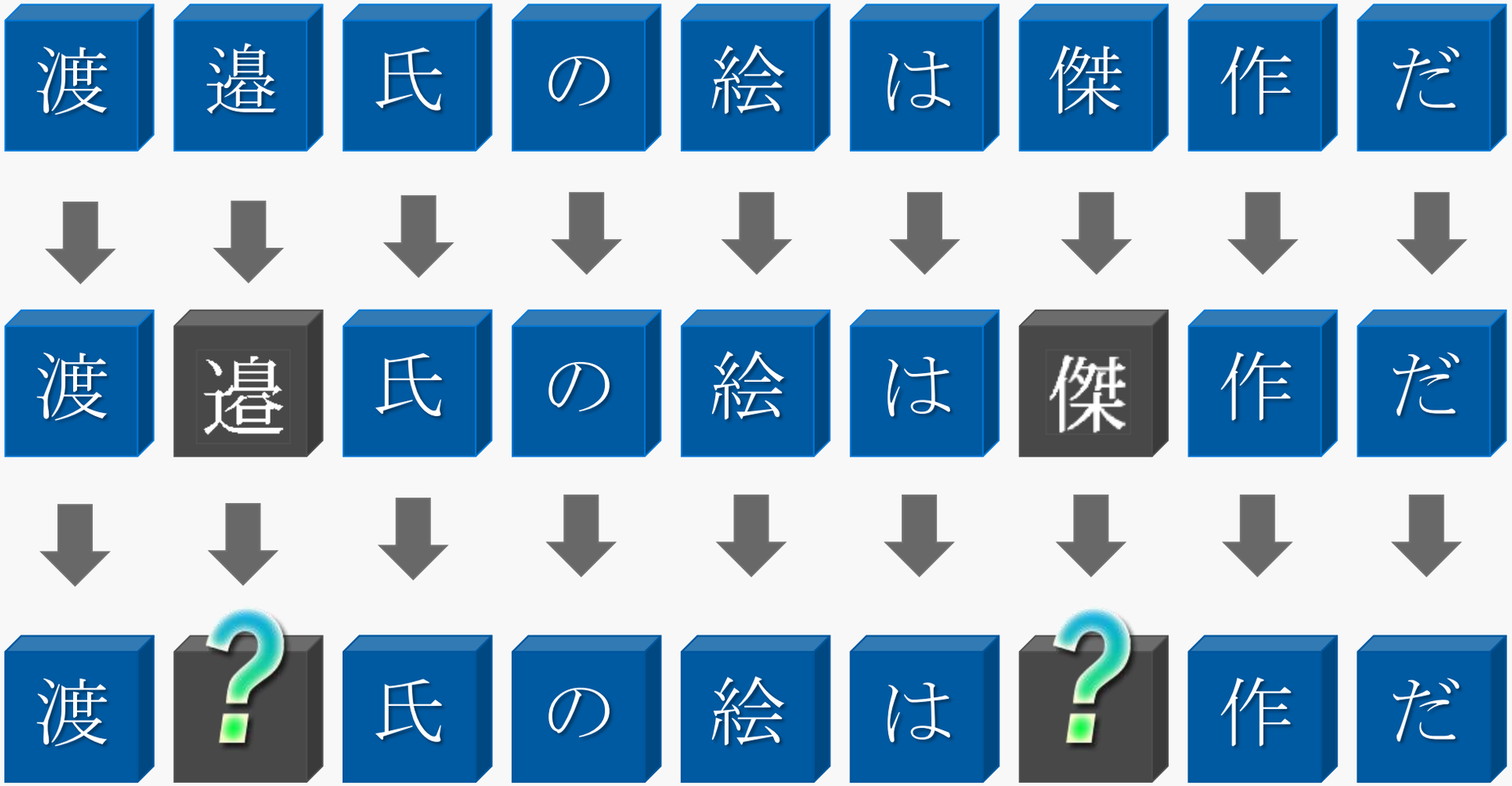
外字を使用する背景・・


無い文字は外字で対応・・・
住民基本台帳ネットワーク統一文字 19,432字
戸籍統一文字 55,267字
登記固有文字 10,330字
登記統一文字 65,597字
相互運用性
一般に・・・
JIS、Unicodeを初め標準に含まれない字
改定常用漢字表
JIS X 0208, JIS X 0213
この他、フォントに含まれず独自に追加した字

外字を必要とする背景とは



BMPの外字使用と外字を使うことの深刻な課題



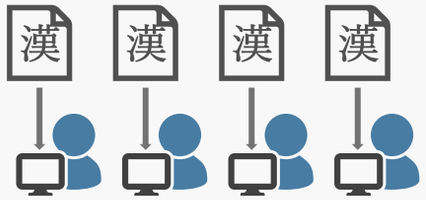
外字を使用した運用

私的コードの割当

1. コードの割り当て、管理
2. グリフの作成
3. フォントファイルへの追加
4. 更新したフォントの配布
5. 他システム連携のためのマッピングテーブルの更新

字体の作成

フォンファイルへの追加



フォントの配布

- 文字のメンテナンス要する
  膨大なコスト
- 外部システムとの相互運用ができない**閉じた文字情報空間**

東北大震災からの教訓

被災自治体

支援自治体

同じ字形であっても外字使用により
コードが異なり、データ連携ができない

Unicode IVS/IVD

Windows 8 の IVSサポート基盤

- **文字情報基盤構築事業で整備された文字、約58,000文字を扱う基盤のサポート**
  - Unicode UTS #37 (IVS/IVD)
  - JIS X 0221:2013 追補版で対応予定
  - 印刷・出版業界で使用されているAJ1-6対応フォントを使用可能に
- **JIS X 0208:1990、JIS X 0213:2000、JIS X 0213:2004の文字全てを同時にサポート**
  - JIS90フォントパック使用することなく、JIS90の字体を
    IMEを使用して入力、表示可能に

グリフデータベース基盤

様々なグリフをIdeographic Variation Databaseとして管理、利用可能にすることで、固有のIDによりグリフの識別を実現

フォントサイズの推移

組み込みなどOSの用途が広がる中で文字、
フォントを増やすことは容易ではない

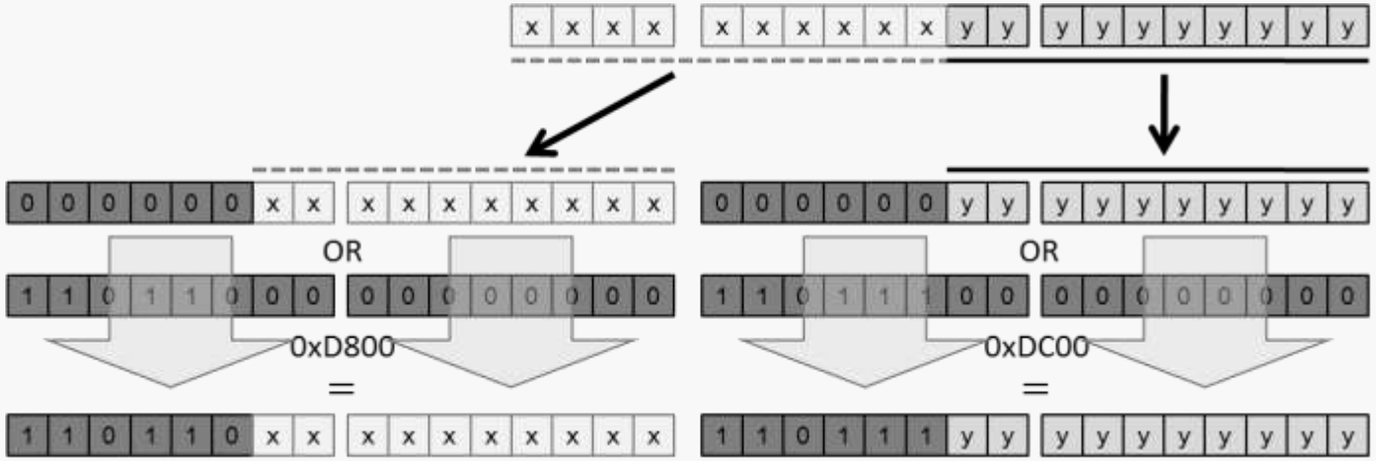
	MSゴシック	MS明朝
Windows 3.1	2,500,692	3,446,204
Windows 2000 SP4	8,272,028	9,135,960
Windows Vista SP2	9,165,480	10,056,872
Windows 7	9,172,500	10,057,108
Windows 8	9,209,540	10,080,360
Windows 10	9,214,692	10,081,800

Windows 8 でIVSに正式対応

- IVS（異体字セレクタ）に対応
  - 文字の入力
  - 文字の出力（表示、印刷など）

１文字＝16ビットではない（UTF-16）

Unicodeのアドレス空間は21ビット
UTF-16は、1文字を16ビットで符号化する方式



レンダリング文字数とデータサイズは一致しない

「	丈	𠮟	叱	」	は	追	加	面	の	文	字
300C	D840 DC0B	D844 DE3B	D842 DF9F	300D	306F	8FFD	52A0	9762	306E	6587	5B57
U+300C	U+2000B	U+2123D	U+20B9F	U+300D	U+306F	U+8FFD	U+52A0	U+9762	U+306E	U+6587	U+5B57

12文字 v.s. 15文字

C#におけるSurrogate Pairの識別

- Char.IsSurrogate(char)
- Char.IsSurrogate(string, int)

C#における見た目上の文字数の取得

- StringInfo.LengthInTextElements

異体字セレクタの仕組み
~Ideographic Variation Sequence~

- 基本となる文字と、その字形を指定する事が可能に
- 見た目の異なる文字であっても、基本となる文字により、意味上の一意性を確保

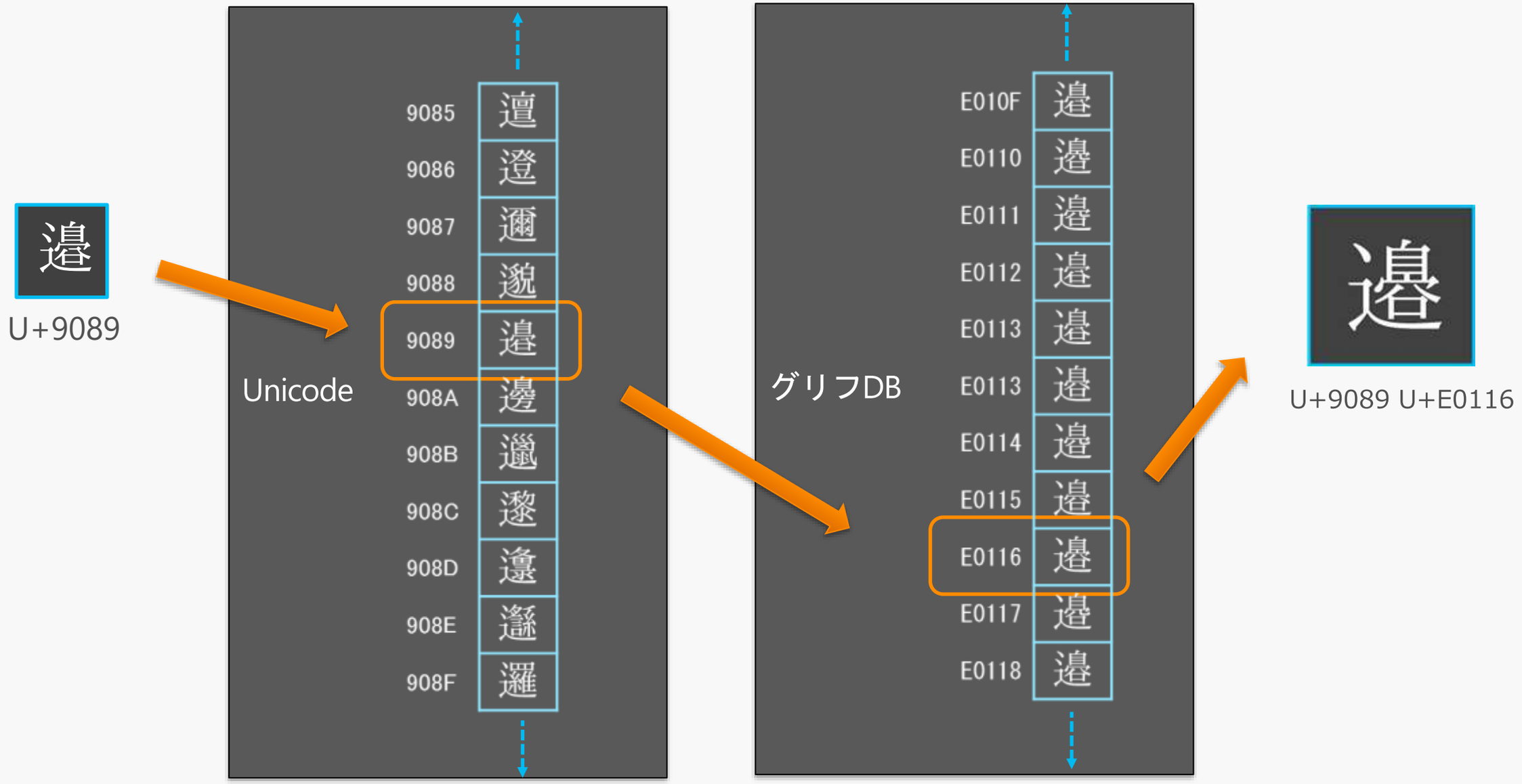
邉
9089

基本となる文字

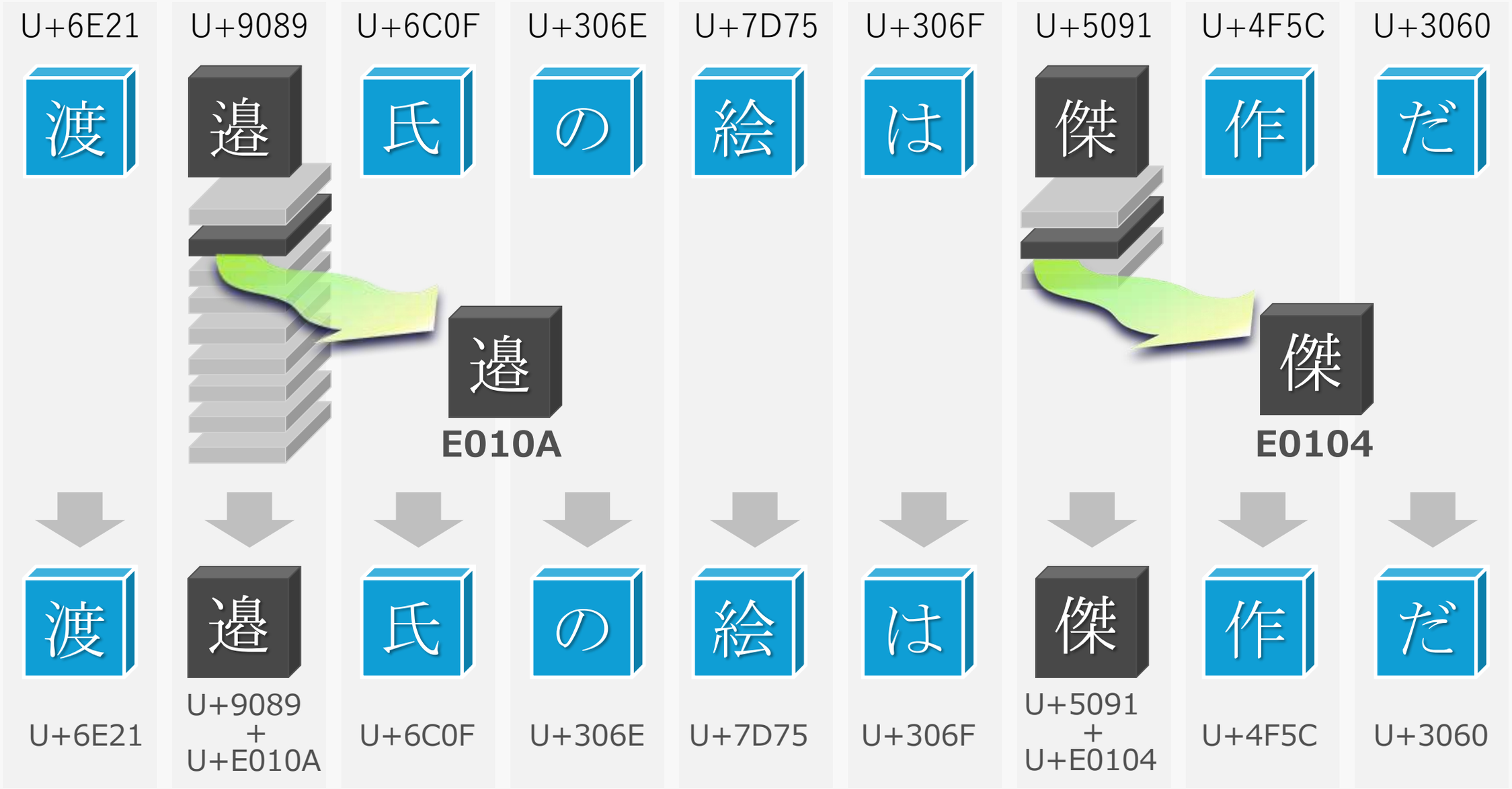
U+9089のバリエーション

- IVD
  - Ideographic Variation Database
- IVS
  - Ideographic Variation Selector
  - Ideographic Variation Sequence

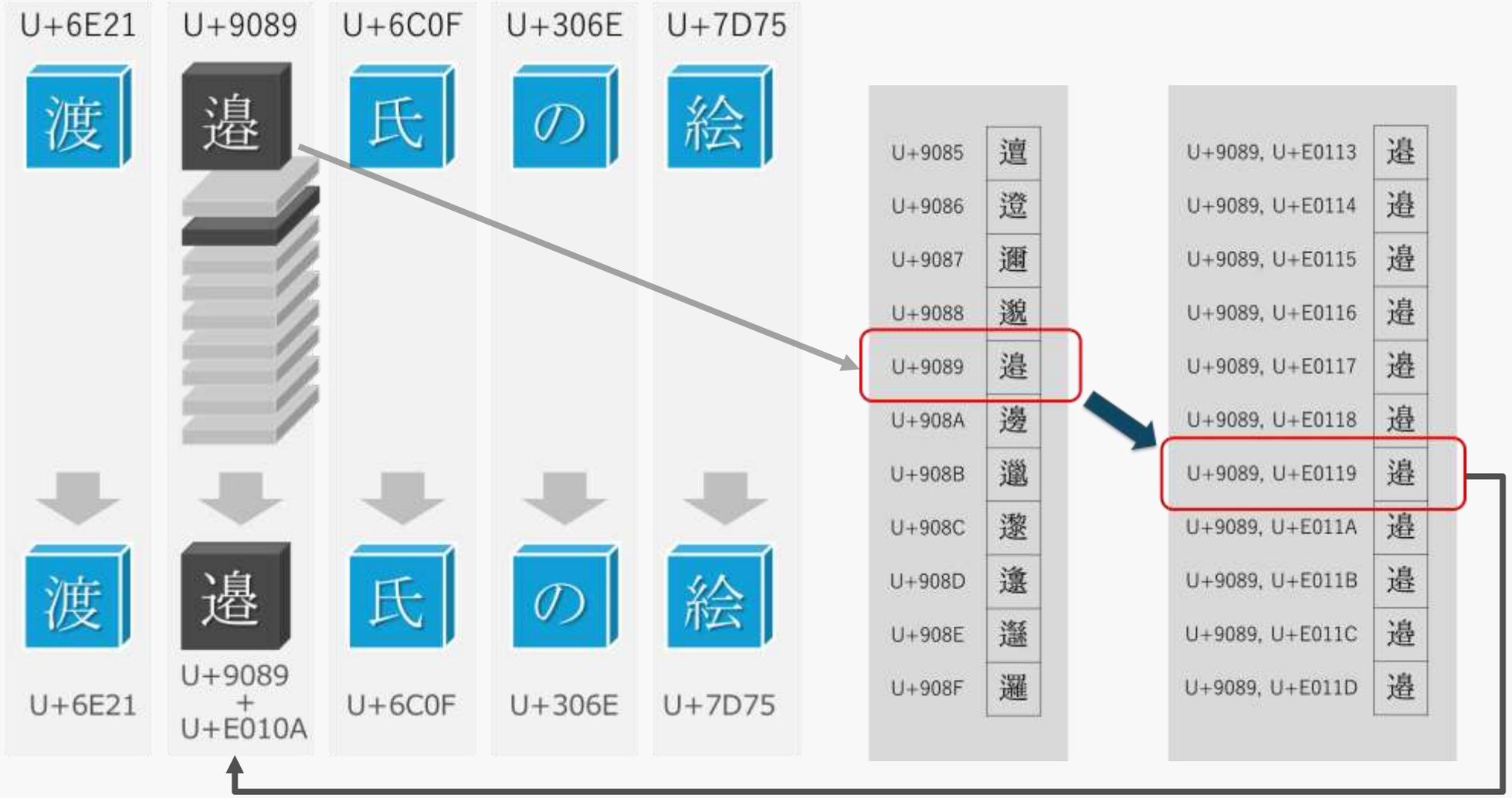
異体字セレクタ


邉
U+9089
Unicode
9085 遖
9086 邆
9087 邇
9088 邈
9089 邉
908A 邊
908B 邋
908C 邌
908D 邍
908E 邎
908F 邏
グリフDB
E010F 邉
E0110 邉
E0111 邉
E0112 邉
E0113 邉
E0113 邉
E0114 邉
E0115 邉
E0116 邉
E0117 邉
E0118 邉
邉
U+9089 U+E0116

グリフの指定が可能に


U+6E21
U+9089
U+6C0F
U+306E
U+7D75
U+306F
U+5091
U+4F5C
U+3060
渡
邉
氏
の
絵
は
傑
作
だ
邉
E010A
傑
E0104
渡
邉
氏
の
絵
は
傑
作
だ
U+6E21
U+9089
+
U+E010A
U+6C0F
U+306E
U+7D75
U+306F
U+5091
+
U+E0104
U+4F5C
U+3060

グリフの指定とIVD


U+6E21
U+9089
U+6C0F
U+306E
U+7D75
渡
邉
氏
の
絵
渡
邉
氏
の
絵
U+6E21
U+9089
+
U+E010A
U+6C0F
U+306E
U+7D75
U+9085 遧
U+9086 遰
U+9087 遷
U+9088 選
U+9089 邉
U+908A 邊
U+908B 邋
U+908C 邌
U+908D 邍
U+908E 邎
U+908F 邏
U+9089, U+E0113 邉
U+9089, U+E0114 邉
U+9089, U+E0115 邉
U+9089, U+E0116 邉
U+9089, U+E0117 邉
U+9089, U+E0118 邉
U+9089, U+E0119 邉
U+9089, U+E011A 邉
U+9089, U+E011B 邉
U+9089, U+E011C 邉
U+9089, U+E011D 邉

16ビットから64ビット (UTF-16の場合)

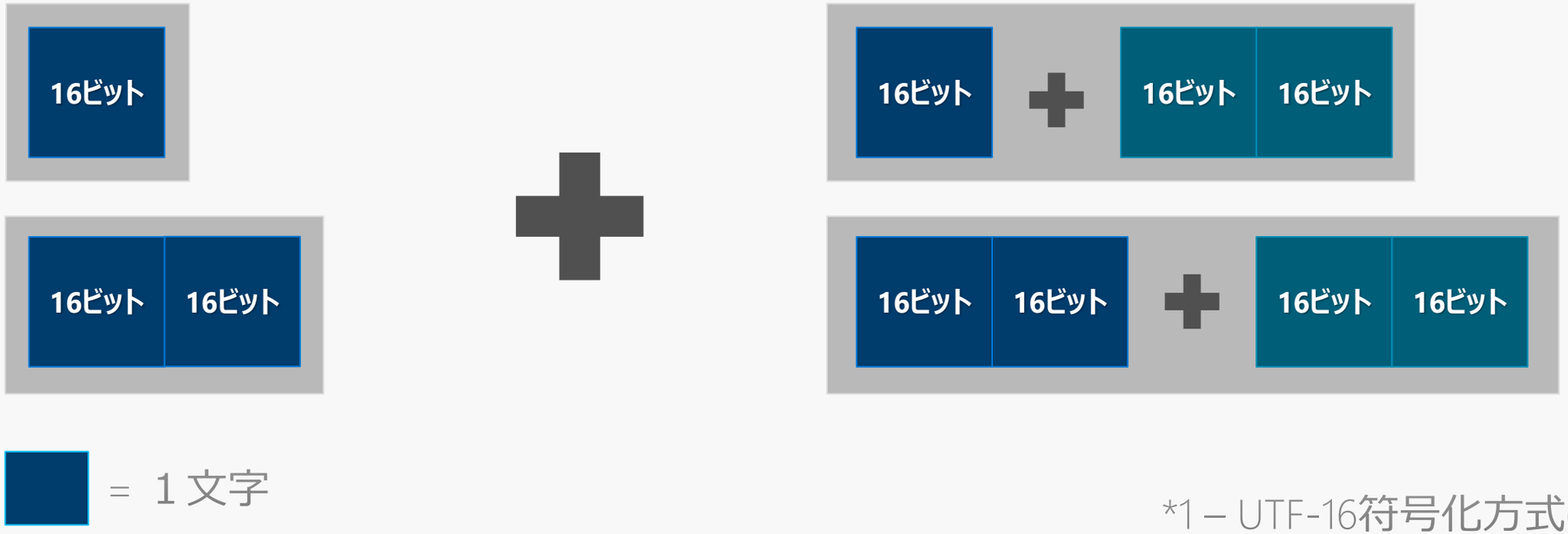
- BMP
  - １６ビット
- サロゲートペア
  - ３２ビット（１６ビット+ １６ビット）

- JIS X 0213、IVSともにサロゲートエリアに割り当てられている

＝ １文字の場合も

実装上の違い

	符号長/１文字	不足する文字への対応
シフトJIS	可変（８、１６ビット）	外字
Unicode 基本多言語面	固定（１６ビット）*1	外字
Unicode サロゲートペア	可変（１６、３２ビット）*1	外字
Unicode IVS/IVD	可変（１６、３２、４８、６４ビット）*1	国際標準



*1 – UTF-16符号化方式の場合

文字によって異なるバイト数（UTF-16）

基本多言語面の文字

U+845B

845B

追加面の文字

U+257A9

D855 DFA9

基本多言語面の文字と異体字セレクタ

U+9F4B U+E0100

9F4B DB40 DD00

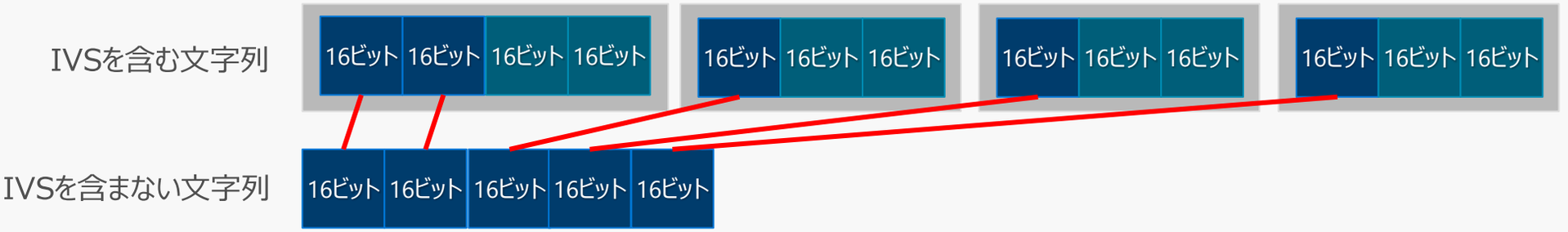
追加面の文字と異体字セレクタ

U+257A9 U+E0100

D855 DFA9 DB40 DD02

各データベースにおける対応状況

- データベースにより異なる対応が異なる文字コード対応
- 照合順序がIVSに対応していないデータベースでは、親文字から構成される文字列による検索ができない。

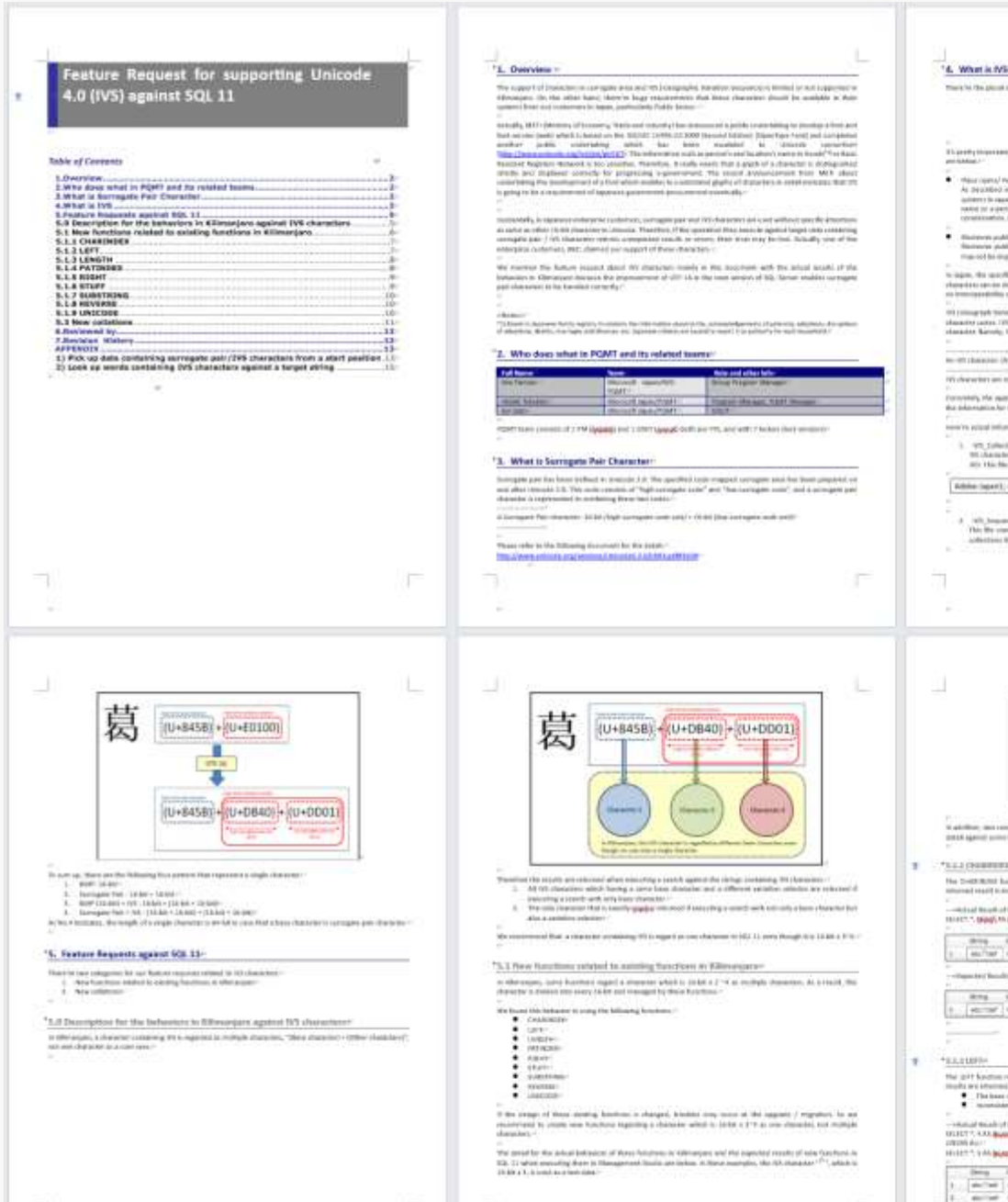


	Upper/Lower	Accent	かな・カナ	Half/Full	IVS	文字情報基盤対応
PostgreSQL 14	citext extension	Unaccent extension				×
MySQL 8.0	CI/CS	AI		M		△
Oracle 21c	ci/cs	ai/as	ks	unicode		△
SQL Server	CI/CS	AI/AS	KS	WS	VSS	◎ *1

※1 - SQL Serverでは、「_VSS」が含まれる照合順序が異体字セレクタに対応

_VS Collation を SQL11で追加

Unicode IVS/IVD 対応 Collation (_VSS)	SQL Server 2017 (14.x)で導入された日本語の照合順序 、**Japanese_Bushu_Kakusu_140** と **Japanese_XJIS_140** で、多様な表意文字のバリエーションセレクターを区別します。バリエーションのシーケンスは、基本文字とバリエーションセレクターで構成されます。この _VSS オプションを選択しない場合、照合順序で異体字の選択が区別されず、比較において異体字の選択が考慮されることはありません。つまり、SQL Server では、並べ替えが同じになるように、バリエーションの選択が異なる同じ基本文字に基づいて構築された文字が考慮されています。詳細については、「Unicode 表意文字のバリエーション データベース」を参照してください。 異体字セレクターを区別する (_VSS) 照合順序は、全文検索インデックスではサポートされていません。全文検索インデックスでは、アクセントを区別する (_AS)、かなを区別する (_KS)、文字幅を区別する (_WS) オプションのみがサポートされます。SQL Server XML と CLR のエンジンでは、(_VSS) 異体字セレクターはサポートされていません。



文字情報基盤

文字情報基盤整備事業

～内閣官房情報通信技術（IT）担当室、経済産業省、IPAにより推進～

平成14～17年度にかけて、戸籍統一文字、住民基本台帳ネットワーク統一文字を対象として、人名、地名等に使用される文字を調査を実施。平成20年度～登記に使用される文字を調査。
情報システムへの実装に向けて、ISO/IEC 10646に必要な文字の追加提案を行う。

これまでの経緯

年		発行主体等	内容
1978	（S53）	日本規格協会	JIS C 6226「情報交換用漢字符号系」
1994	（H6）	法務省	戸籍事務の電算化開始
1997	（H9）	閣議決定	「行政情報化推進基本計画（改定）」JIS3,4水準制定により解消。残る外字の交換ルール策定。
2000	（H12）	日本規格協会	JIS X0213「情報交換用符号化拡張漢字集合」（JIS第3,4水準）
2002	（H14）	総務省	住民基本台帳ネットワーク統一文字
2002	（H14）	経済産業省他	汎用電子情報交換環境整備プログラム開始（戸籍、住民基本台帳ネットワーク、登記を検証）
2003	（H15）	IT戦略本部決定	「e-Japan 戦略Ⅱ」公開用文字情報データベースの構築。文字コード規格を整備。
2004	（H16）	法務省	戸籍統一文字（戸籍手続オンラインシステムの構築のための標準仕様書）
2005	（H17）	IT戦略本部決定	「IT政策パッケージ-2005」
2008	（H20）	IT戦略本部決定	「IT政策ロードマップ」
2010	（H22）	経済産業省、IPA	文字情報基盤プロジェクト開始

これまでの経緯

年		発行主体等	内容
2010	（H22）	経済産業省、IPA	文字情報基盤プロジェクト開始
2011	（H23）	IT戦略本部決定	「電子行政推進に関する事本方針」文字情報基盤の活用
2011	（H23）	経済産業省、IPA	**文字情報基盤プロジェクト成果公開**　IPAmj 明朝フォント、文字情報対応表
2013	（H25）	閣議決定	**「世界最先端IT国家創造宣言」**文字情報基盤を原則
2014	（H26）	総務省	「**電子自治体の取組を加速するための10の指針**」
2014	（H26）	CIO連絡会議	「**電子行政分野におけるオープンな利用環境整備に向けたアクションプラン**」 導入ガイドの活用、縮退マップの整備

世界最先端IT国家創造宣言
内閣閣議決定

従来政府が担っていたサービスの提供機能を民間にも開放し、官民の協働によって、より利便性の高い公共サービスを創造する。国民がステークホルダーとして積極的に参加できるよう、クラウドを活用したオープンな利用環境を、データ・フォーマット、用語、コード、文字等の標準化・共通化、アプリケーション・インターフェイス（API）の公開等を行いつつ整備する。特に文字の標準化・共通化に関しては、今後整備する情報システムにおいては、国際標準に適合した**文字情報基盤**を活用することを原則とする。

電子行政分野におけるオープンな利用環境整備に向けたアクションプラン

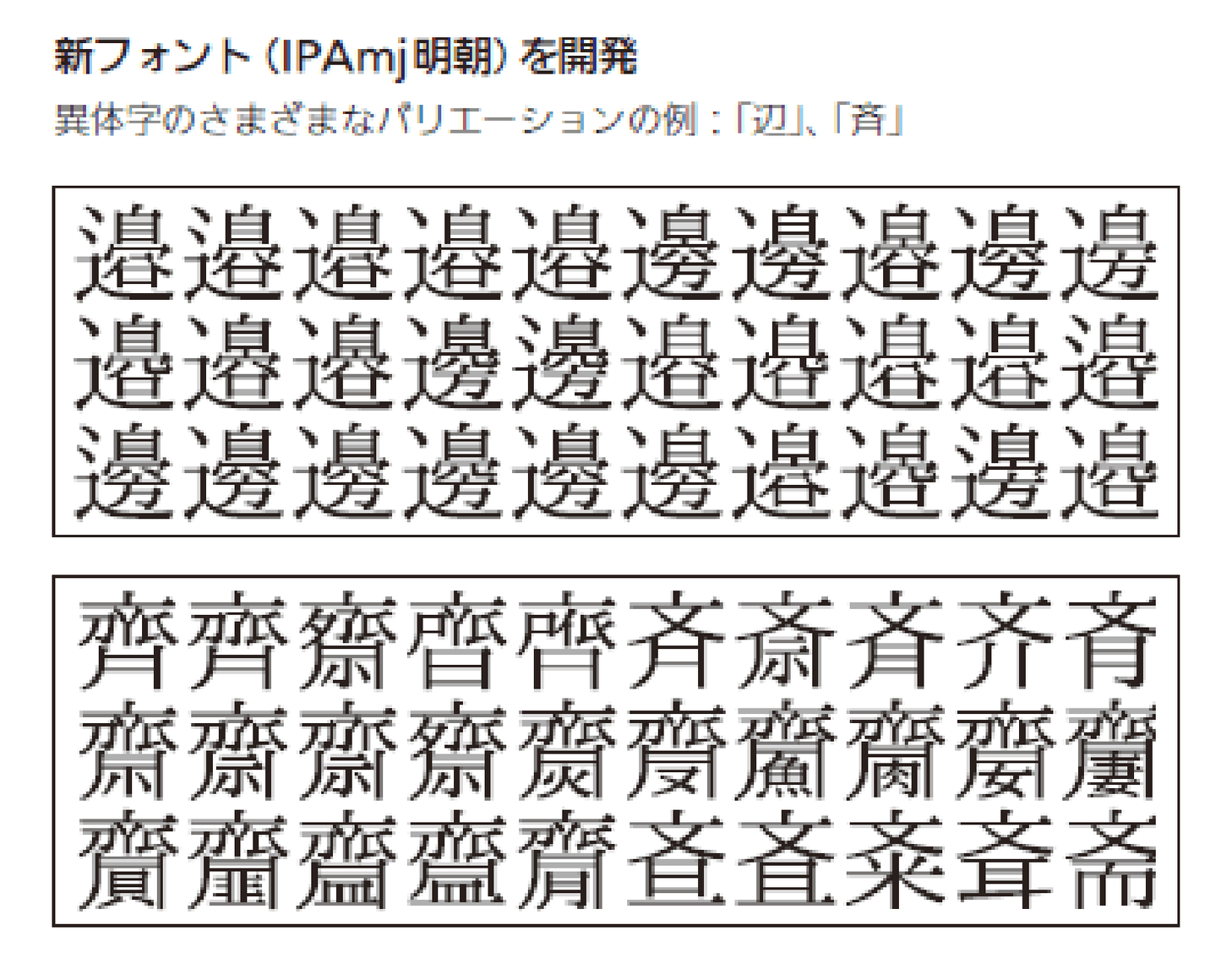
各府省情報化統括責任者（CIO）連絡会議決定

今後整備する政府情報システムにおいては、国際標準に適合した**文字情報基盤**を活用することを原則とする。

各府省は、経済産業省を中心に策定した「文字情報基盤導入ガイド」を参照しつつ、導入を推進する。また、経済産業省は、各府省の円滑な導入を支援するため、文字情報基盤の文字（約６万文字）を、市販コンピュータで特別な設定無しで活用できるJIS 範囲の文字への変換を行う際に参照する縮退マップの提供を、平成26 年度中に実施する。

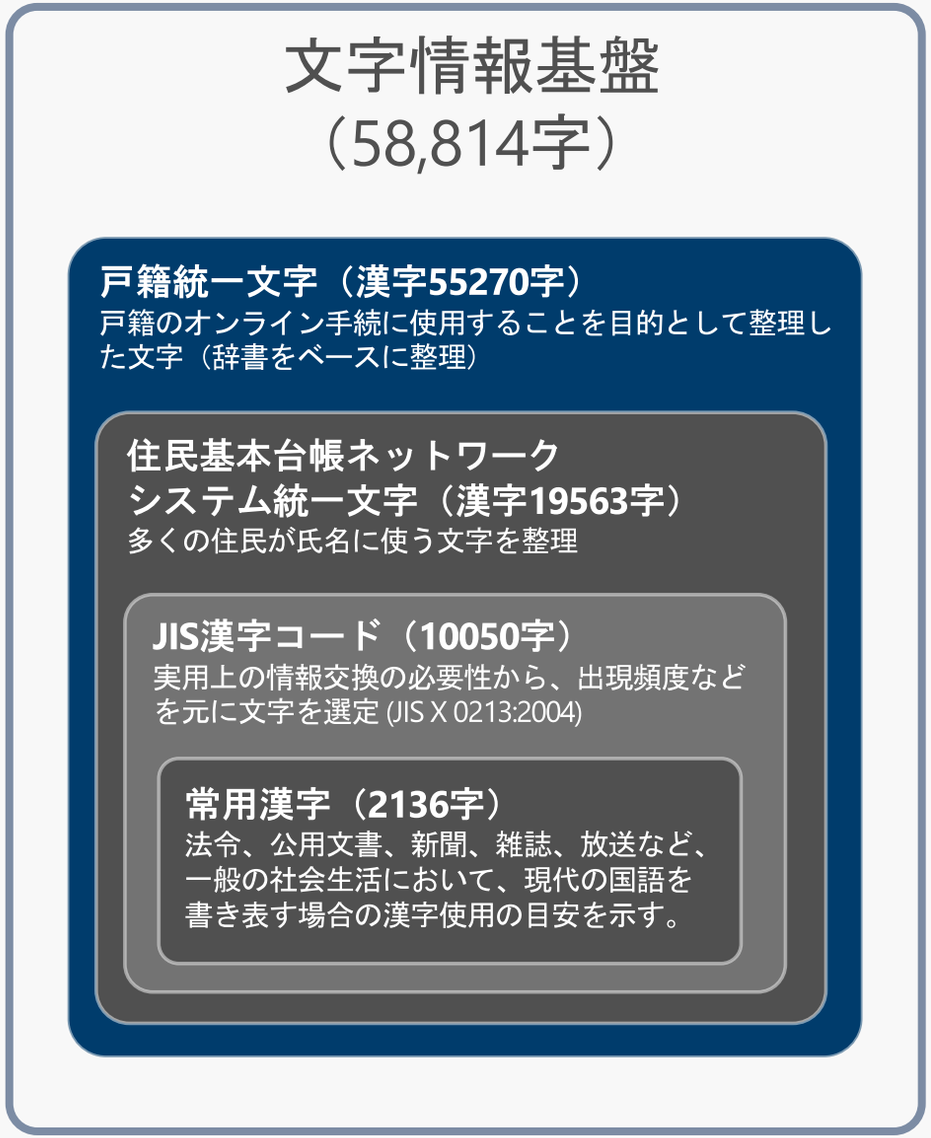
IPAmj明朝

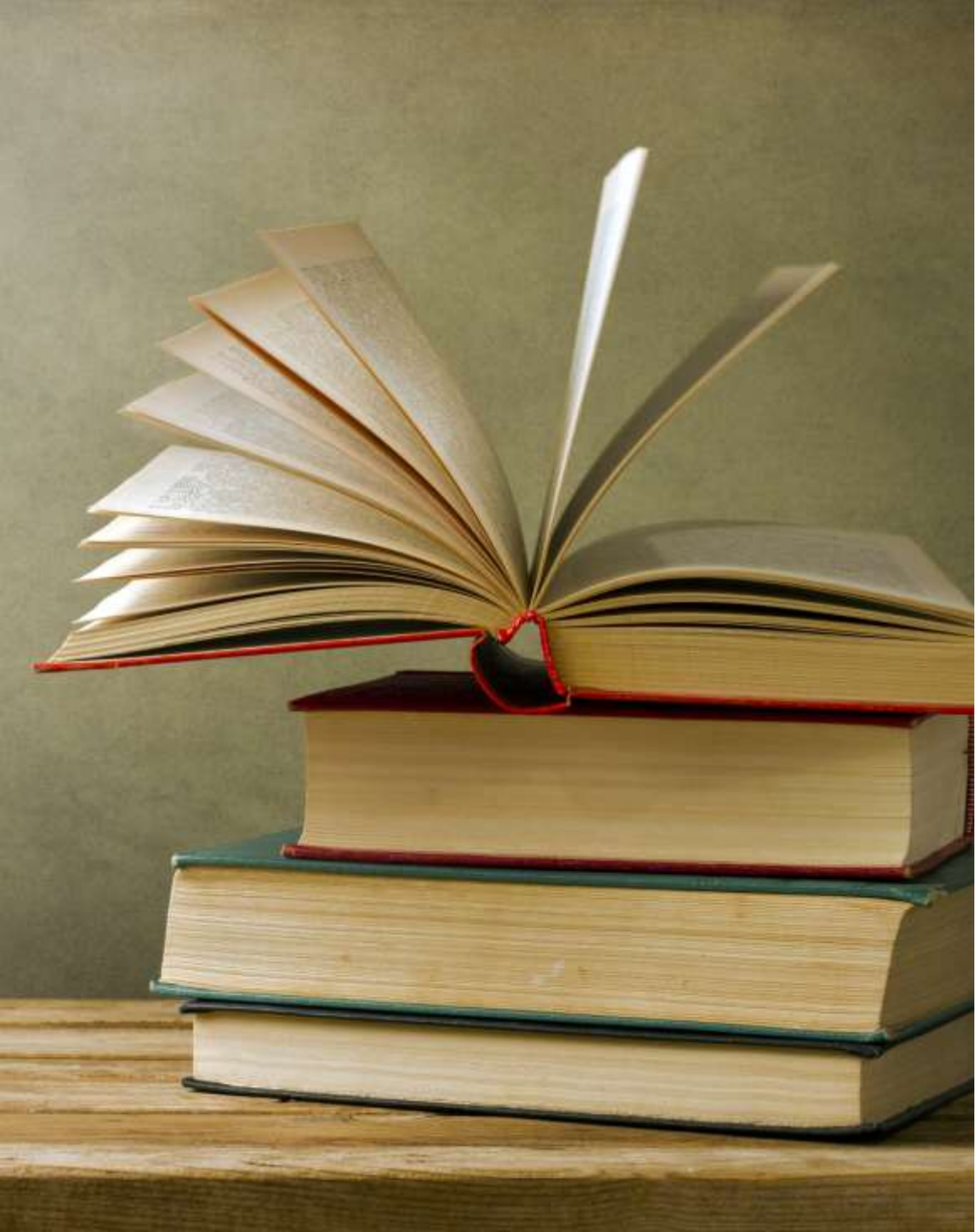
初版：2011年10月



文字情報基盤概要

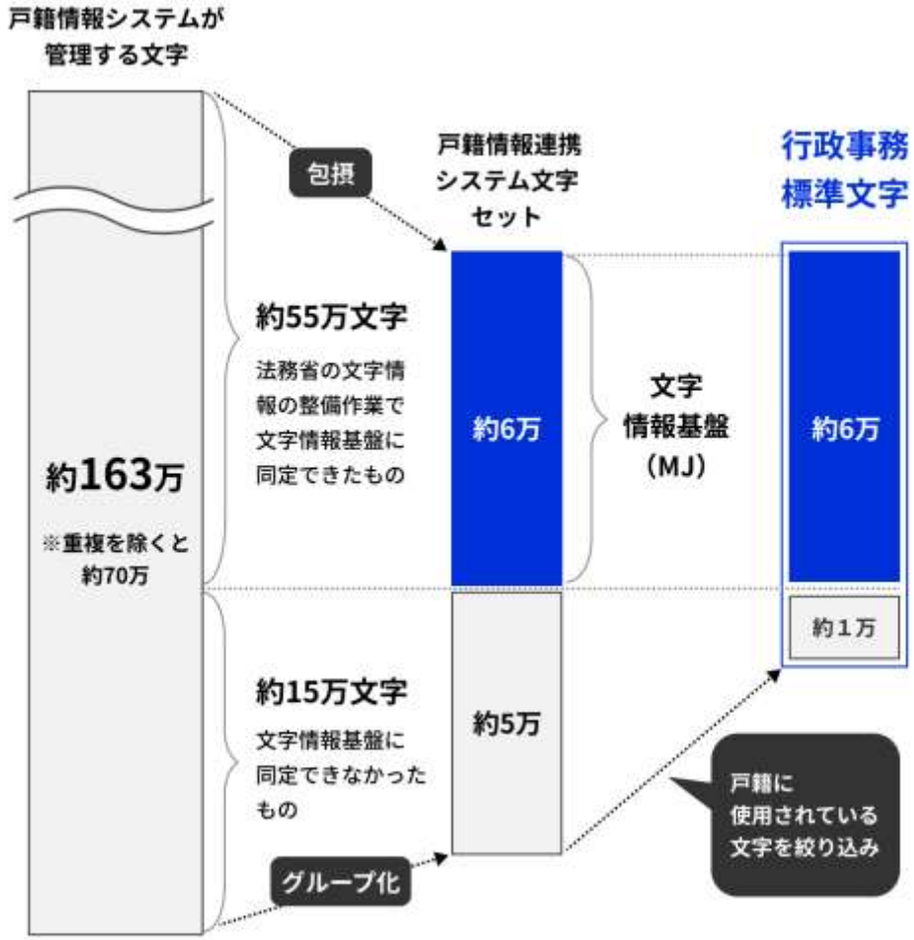
- 我が国で行政業務上必要とされる人名漢字等約6万文字を、国際標準に則り提供
  - IPAmj明朝フォント
  - 文字情報一覧表
  - http://mojikiban.ipa.go.jp/





行政事務標準文字

行政文字標準文字とは



https://www.digital.go.jp/policies/local_governments/character-specification

外字による様々な課題の解消

- **外字作成コスト**
  - 地方公共団体ごとや、一つの地方公共団体の中でも手続を行う部署ごとに文字データを作成・維持しています。
- **地方公共団体の職員・住民の負担大**
  - 転入時に住民票の写しを即日発行できません。地方公共団体の職員が外字を作成したのち、住民は後日住民票の写しを受領します。
  - 地方公共団体等の公的機関では、手続きの際、氏名の入力誤りが発生しないようにするための確認作業等で大きな負担が発生します。
  - 地方公共団体ごとにコンピューターで管理する文字が異なることで、効率的な行政サービスの実施および大規模な災害発生時の迅速な対応等の妨げとなります。
- **システム選択時の制約**
  - システム更新の際には、外字が支障となって、同じベンダーに依存し続けるベンダーロックインの状態になりやすく、文字の作成や維持・管理にコストがかかる要因にもなります。
- **システム間での情報連携を阻害**
  - 住民サービスに用いられているコンピューターのデータが別々のシステムで管理されているため、外字を使用している場合は、異なる地方公共団体間のデータ連携だけでなく、同じ地方公共団体の中の組織であっても、部署が異なればシステム間のデータ連携の際に文字化け等が発生し、情報の連携を阻害する要因になっています。

https://www.digital.go.jp/policies/local_governments/character-specification

一般社団法人 文字情報技術促進協議会

https://moji.or.jp

全ての文字を全ての環境へ

- フォント技術
- 標準化
- 技術支援
- 政策

※Unicode Consortium alliance organization

https://moji.or.jp

文字が描く、デジタル社会の未来図

～ 一般社団法人文字情報技術促進協議会 15周年シンポジウム ～

文字・フォントの最新動向から政府の外字問題解消の取り組みまで、文字・フォントに関する今をご紹介

日時： **12月12日（金）**
13:00 ～ 17:00

場所： ＪＰタワーホール＆カンファレンス
東京都千代田区丸の内二丁目7番2号
KITTE 4, 5階

参加費： 無料

内容	講演者
開会あいさつ	一般社団法人文字情報技術促進協議会 会長 小林 龍生
基調講演（1）	デジある庁 デジタル社会共通機能グループ 統括官 楠 正憲
基調講演（2）	Unicode Consortium Technical Director Ken Lunde
セッション1	大学共同利用機関法人　人間文化研究機構 国立国語研究所 教授 高田 智和
セッション2	一般社団法人文字情報技術促進協議会 理事 / 日本電気株式会社 袴田 博之
セッション3	Monotype 開発部 津田 昭 モリサワデ ザイン部門長 岡 繁樹 株式会社イワタ 代表取締役社長　水野昭
パネルディスカッション	
閉会あいさつ	一般社団法人文字情報技術促進協議会 事務局長 田丸 健三郎

文字情報技術促進協議会

Microsoft