パネルディスカッション

日本のITと行政との接面での《外字》とは何か

日本マイクロソフト株式会社 業務執行役員 NTO
情報処理学会 情報規格調査会（ISO/IEC JTC 1 国内審議団体） 規格役員
ISO/IEC JTC1 SC2委員（Coded-Character Set）
一般社団法人 文字情報技術促進協議会 理事・事務局長

田丸 健三郎

読み手と書き手の字形の一致を目指して

- 一般社団法人 文字情報技術促進協議会は、2010年に「**読み手と書き手の字形が一致する文字情報交換の実現**」を目指してIVS普及促進協議会」として発足。（2019年社団法人に改組）

外字の無い世界の実現

- 情報報を残せる
- 交換できる
- 読み手と書き手が同じ字形を前提にできる
- 実装コストの削減

Operating Values

- Local solutions require global collaboration
- Localization respects and empowers users
- Interoperability across platforms serves you – and the greater good
- Transparency and open source ensure: Reliability — Security — Stability

https://home.unicode.org/about-unicode/

文字情報技術促進協議会設立時説明資料より抜粋

メーカー独自拡張（外字）による混乱の歴史

- マイクロソフト社が開発した日本語版MS-DOSがサポートしたシフトJISを**OEMメーカーが独自拡張した**ことにより、システム間の**相互運用が問題に**なる。[～ 1992]
  - 他システムとデータ交換・共有を行わない閉じたシステムの想定であっても、後にデータ共有・交換が必要となる事例は枚挙に遑がない。

閉じたシステムで使用することを前提としても、対象システム外での同じ文字利用ニーズは必ず生じ、結果として外字が当初予定していたシステム以外でも使用される状況を生む。

メーカーにより拡張された文字を搭載
NEC PC-9800シリーズ
IBM PS/55 シリーズ
富士通 FMRシリーズ

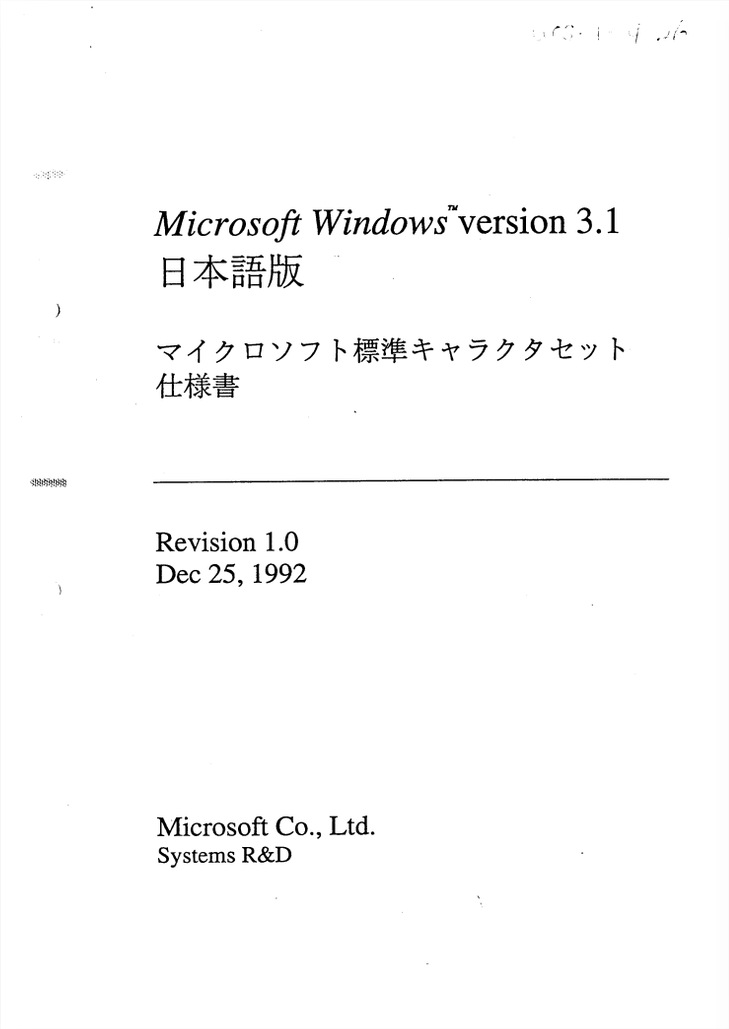
過去にメーカー拡張された文字を使用して作成されたデータをそのまま扱える標準システムは存在しない。（情報の喪失）

相互運用性の喪失

各ベンダーが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。

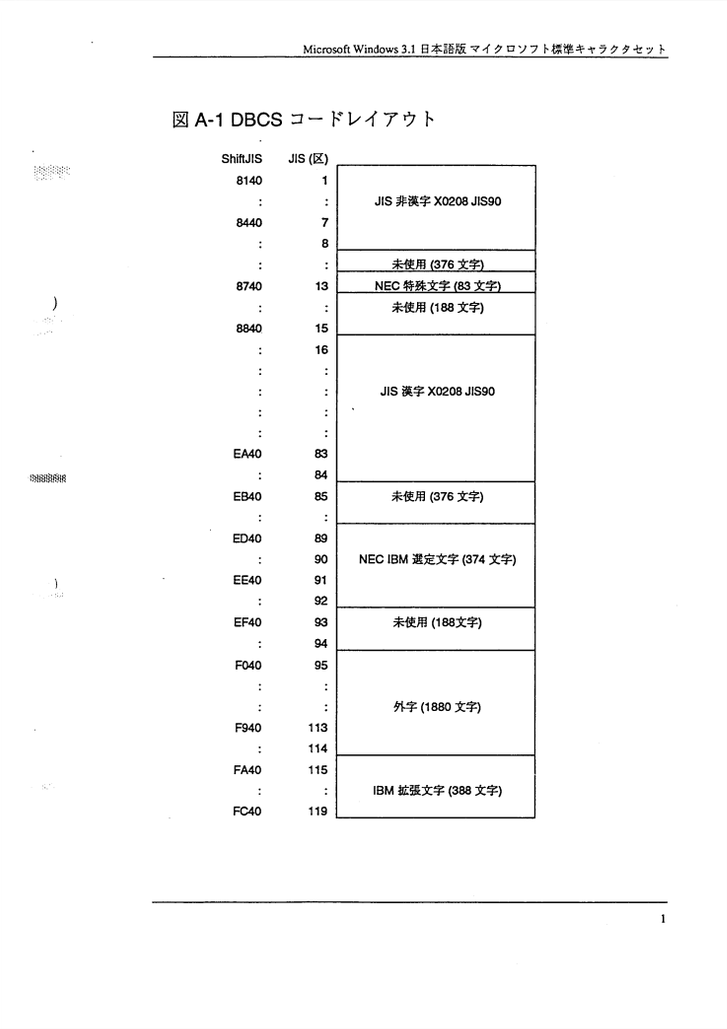


Microsoft Windows™ version 3.1
日本語版

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



Microsoft Windows 3.1 日本語版 マイクロソフト標準キャラクタセット

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	

1

セキュリティにおける標準的な考え方として

クローズドな環境は存在しない

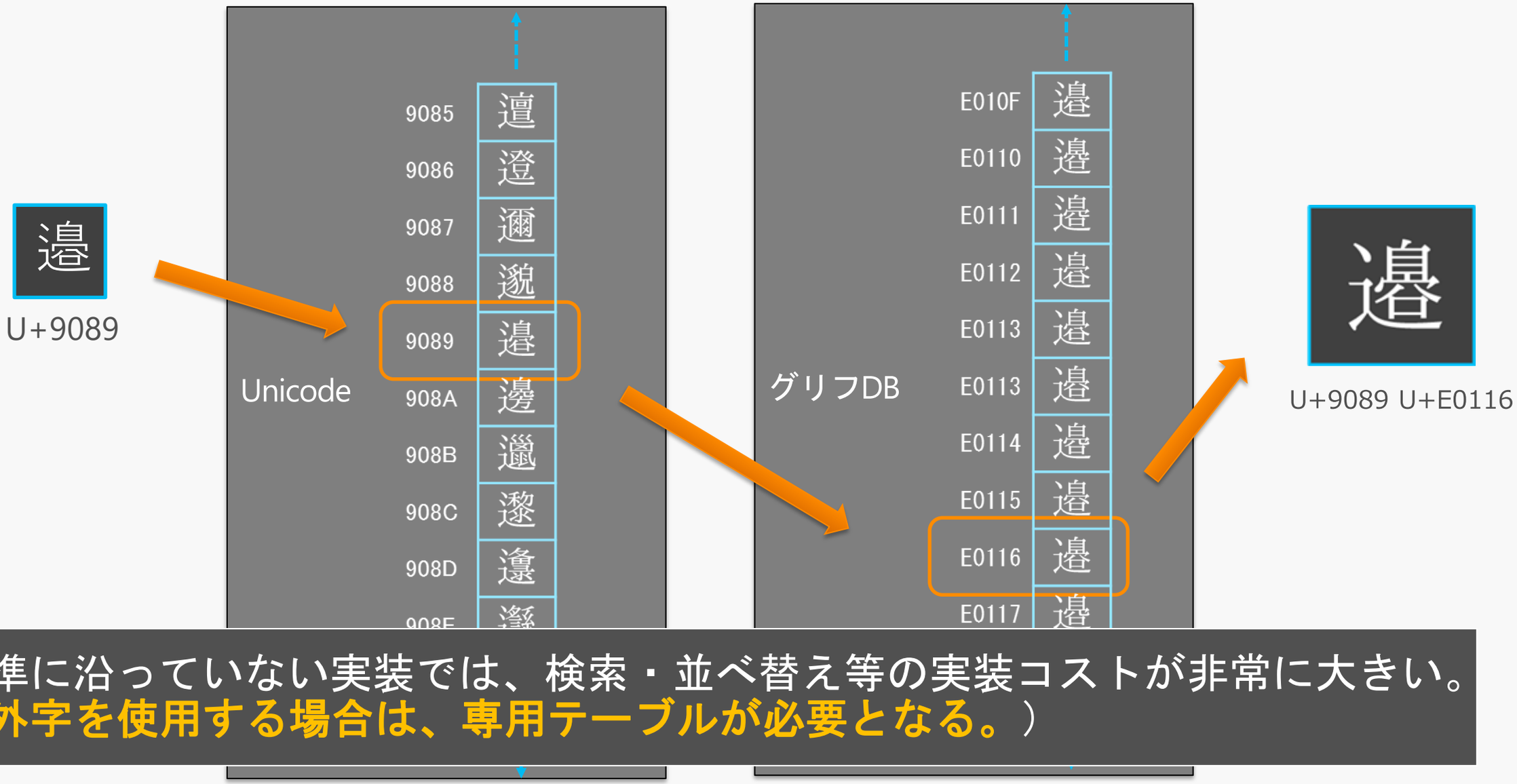
- ユーザーは入力・表示されている文字のコード、データに含まれる文字種など一切意識しない。
- 固有な文字は入力が困難なことが多く、コピー・ペーストによる文字の流通は自然と発生する。
- データ活用の重要性が増すなか、データアーカイブ上の大きな問題となるだけでなく、データ活用における阻害要因。

クローズド環境であれば外字使用は問題ないというのは情報のアーカイブ、文字情報の汚染など非常に多くの問題を抱えている。

自治体における外字の問題

- 災害時に標準的なシステムが被災自治体のデータを扱えない。
  （東北大震災では大きな問題に）
- 様々なアウトソースコストを増大させる
  - 標準化されていない文字は、専用の環境を必要とし外部事業者への委託業務のコストを増大させる。
- 既に外字を利用している
  - 新たな外字は、各自治体独自に外字集合との対応関係表を作成しなくてはならず、多大なコスト増となる。（独自外字環境を持つ自治体 x マッピング作業）

異体字の符号化は標準化されている



標準に沿っていない実装では、検索・並べ替え等の実装コストが非常に大きい。
（**外字を使用する場合は、専用テーブルが必要となる。**）

標準化された文字を対象とする共通実装

- 文字情報の処理は、標準化され共通実装として各ベンダーで共有されることにより、情報処理における相互運用・相互接続時のトラブルの最小化を実現している。

International Components for Unicode（ICU）は、ほぼ全ての主要ITベンダーにより使用され共同でメンテナンスされている。（例：Apple, Microsoft, IBM, Google, Meta, Oracle, Adobe, Salesforce）
ICUは、主に以下の機能を提供している。（標準にない外字は対象外）

- 文字・文字列比較
- 正規表現
- 正規化
- 日付・時間フォーマット
- 照合順序
- など

ICU - International Components for Unicode

https://icu.unicode.org

情報システムの基本である文字コード

日本は、先進国で文字コードが
未だ安定しない唯一の国であることを
どのように考えるか