UnicodeとJIS X 0213
～情報システムにおける日本語処理～

田丸 健三郎
マイクロソフト イノベーション センター
http://www.microsoft.com/japan/mic

アジェンダ

- プラットフォームと相互運用性
- 文字集合の重要性
- 文字コードの歴史とマイクロソフト
- 文字コード…ちょっとその前に
- JIS?
- Unicode

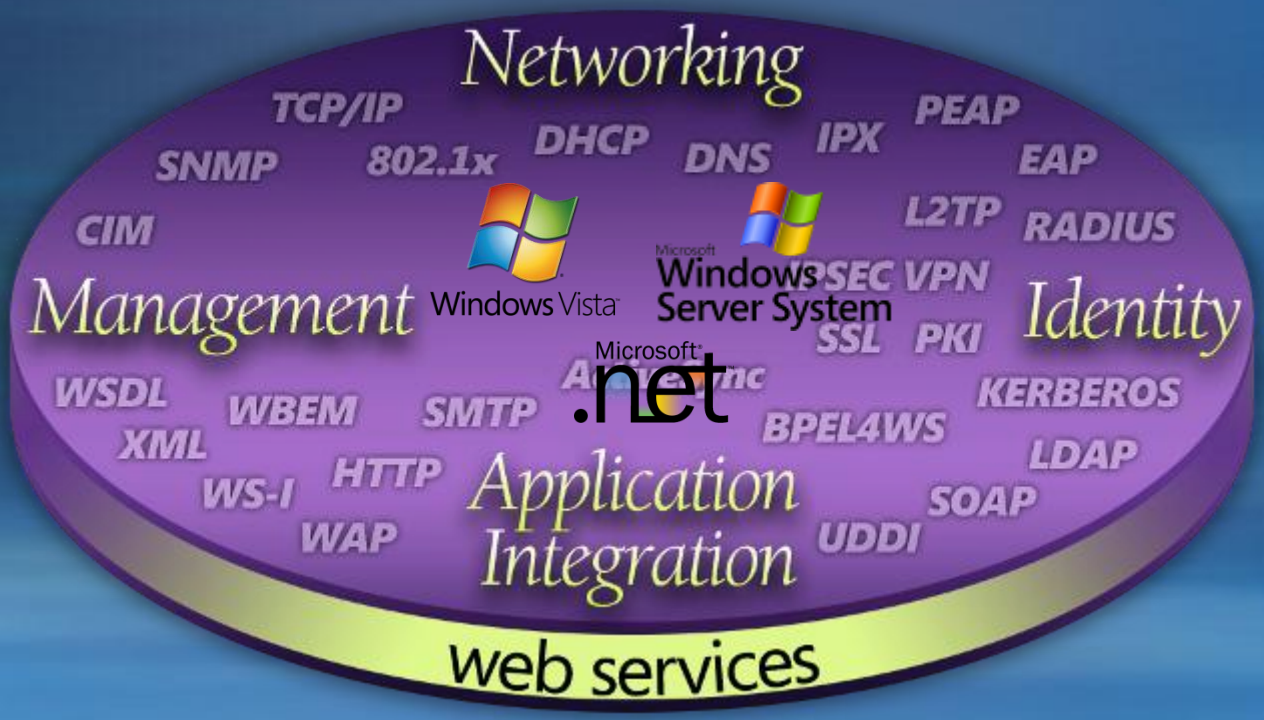
用語

- Shift JIS
- Unicode
- UCS (Universal Character Set)
- BMP (Basic Multilingual Plane)
- Surrogate Pair
- 8bit, 16bit, 32bit
- Wide Character
- 半角、全角
- UI (User Interface)

- JIS78 – JIS C 6226:1978
- JIS83 – JIS C 6226:1983
- JIS90 – JIS X 0208:1990
- JIS97 – JIS X 0208:1997
- JIS2000 – JIS X 0213:2000
- JIS2004 – JIS X 0213:2004
- 補助漢字 – JIS X 0212:1990
- 10646 – ISO/IEC 10646
- ASCII – ANSI INCITS 4

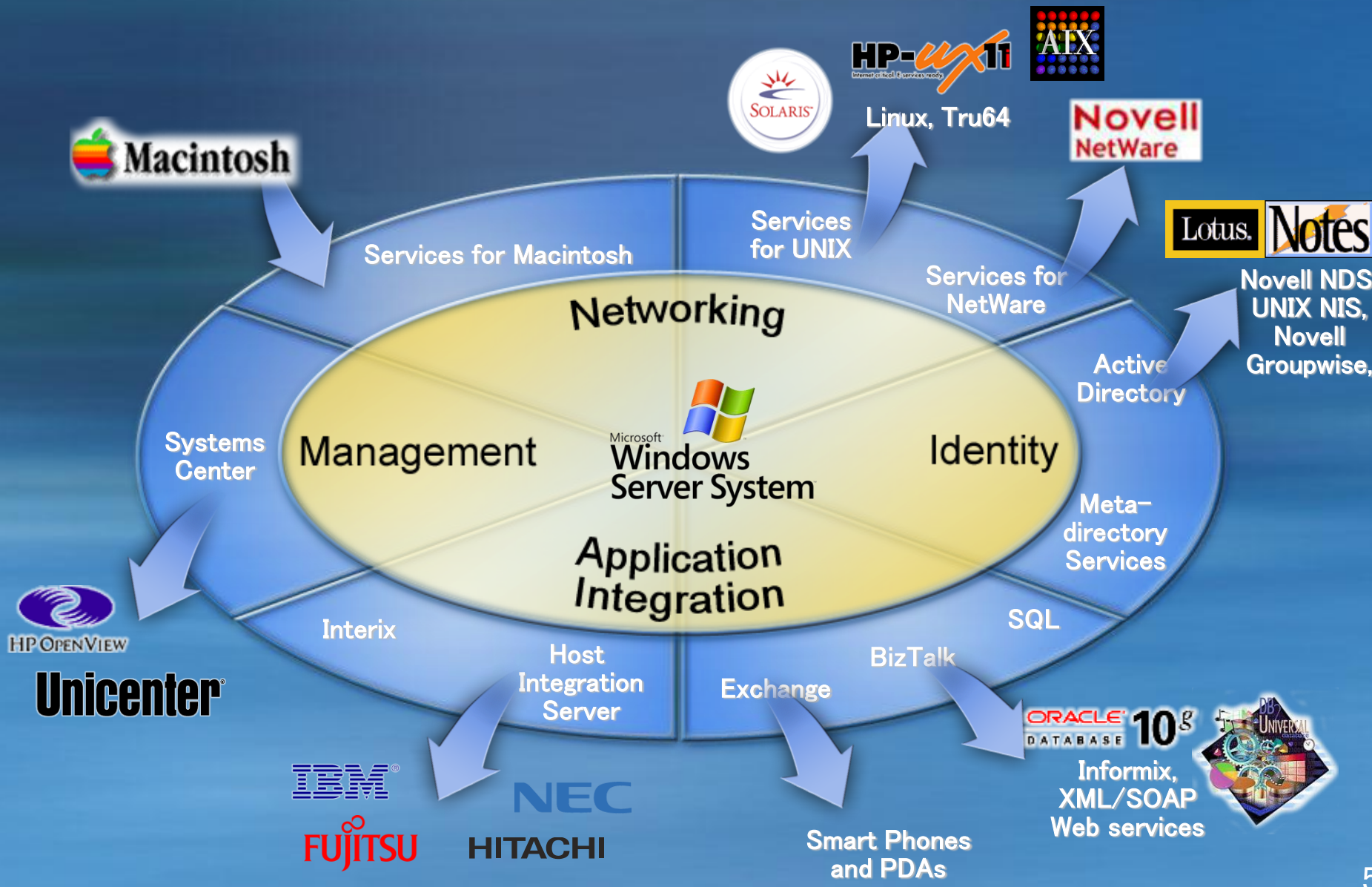
注意：略称は他の文献などで使われている例と違うことがあります。例えば、JIS78 vs. 78JIS

標準に基づくオープンなプラットフォーム



www.w3c.org www.ieee.org http://www.ecma-international.org/ www.ietf.org www.ws-i.org

標準に基づく連携による相互運用性

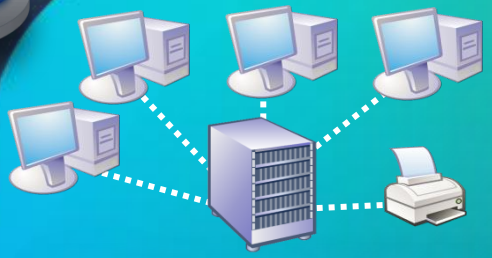


符号化文字集合の重要性

安定した符号化文字集合は、ICT機器間の相互運用性確保の根幹となる基本要件の一つ

Seamless Computing

Web



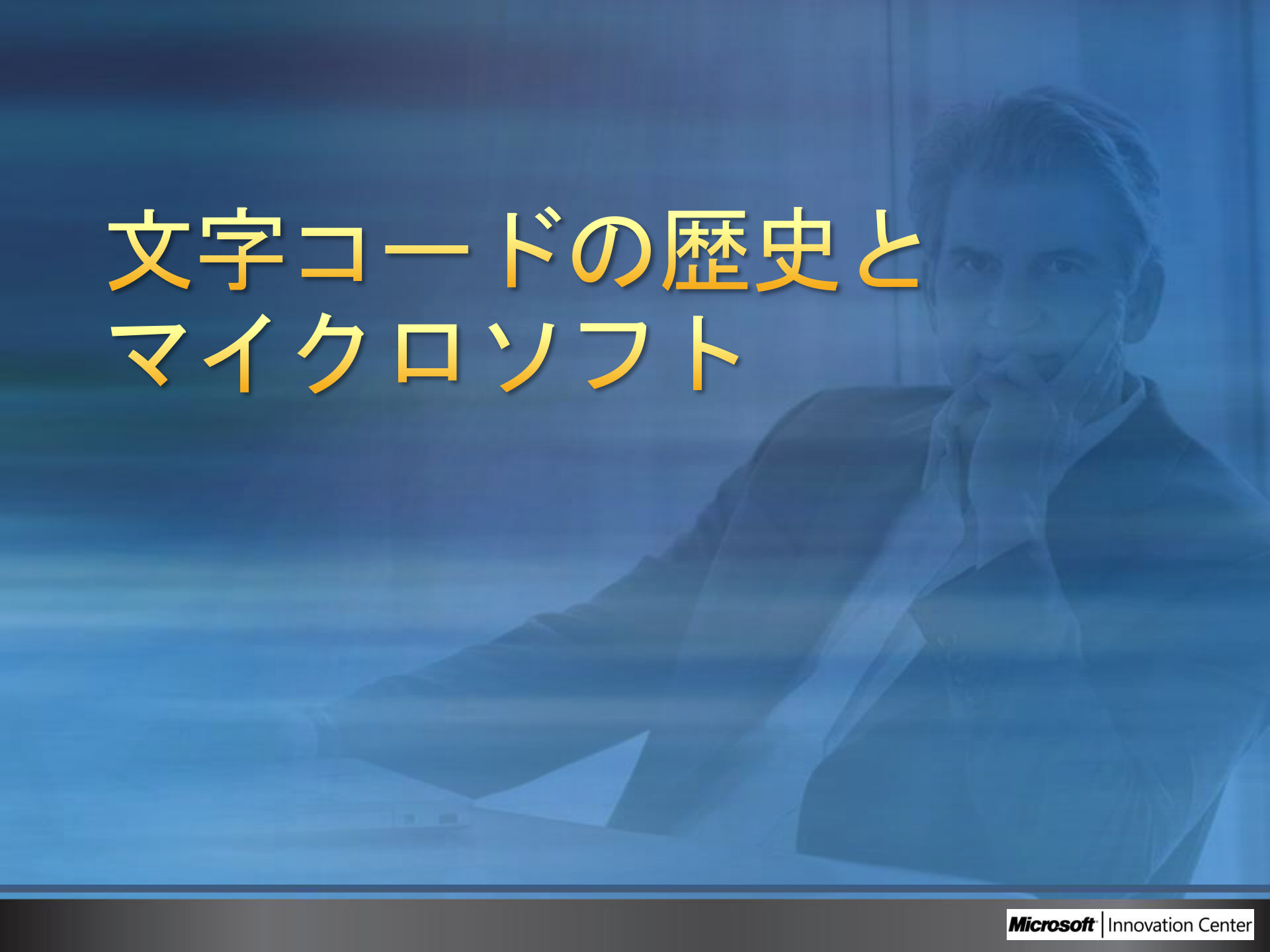
Client-Server

JIS

PC

ISO JTC 1 IEC
INFORMATION TECHNOLOGY STANDARDS

文字コードの歴史とマイクロソフト



日本語文字コード標準とMicrosoft のOS

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 2008, Windows XP (with JIS2004 Pack)

第一世代

日本語対応のソフト開発を容易に ~ Shift JIS 誕生 ~

- 1982年12月MS-DOS 2.0説明会にて発表
- 1983年 MS-DOS 2.0 で対応
- 半角は 1Byte、全角は2Bytes
- 1st Byteで半角と全角を判断
- 2nd Byte に、Delimiter(‘¥’ など)のコードが存在
- 文字集合は JIS78(または、JIS83)を基本
- メーカー拡張文字を収容可能
- 拡張性に乏しい

第二世代

符号化方式と文字集合を定義
~マイクロソフト標準キャラクタ セット~

- 1992年12月 Version 1.0 完成
- 1993年5月 Windows 3.1 で対応
- 符号化方式は、シフト JIS
- 文字集合
  - 最新規格 (JIS90)を採用する
  - 各メーカーの拡張文字を「できるだけ」含み、重複も含む
  - すべての文字を 10646とマッピング
  - 外字領域を残す
- 「マイクロソフト標準キャラクタ セット」の仕様を凍結
  - 将来の改変、拡張はしない
  - 将来の文字拡張は、UCSのみ

第三世代
UCS でのみ利用可能な文字拡張
~ 補助漢字の追加 ~

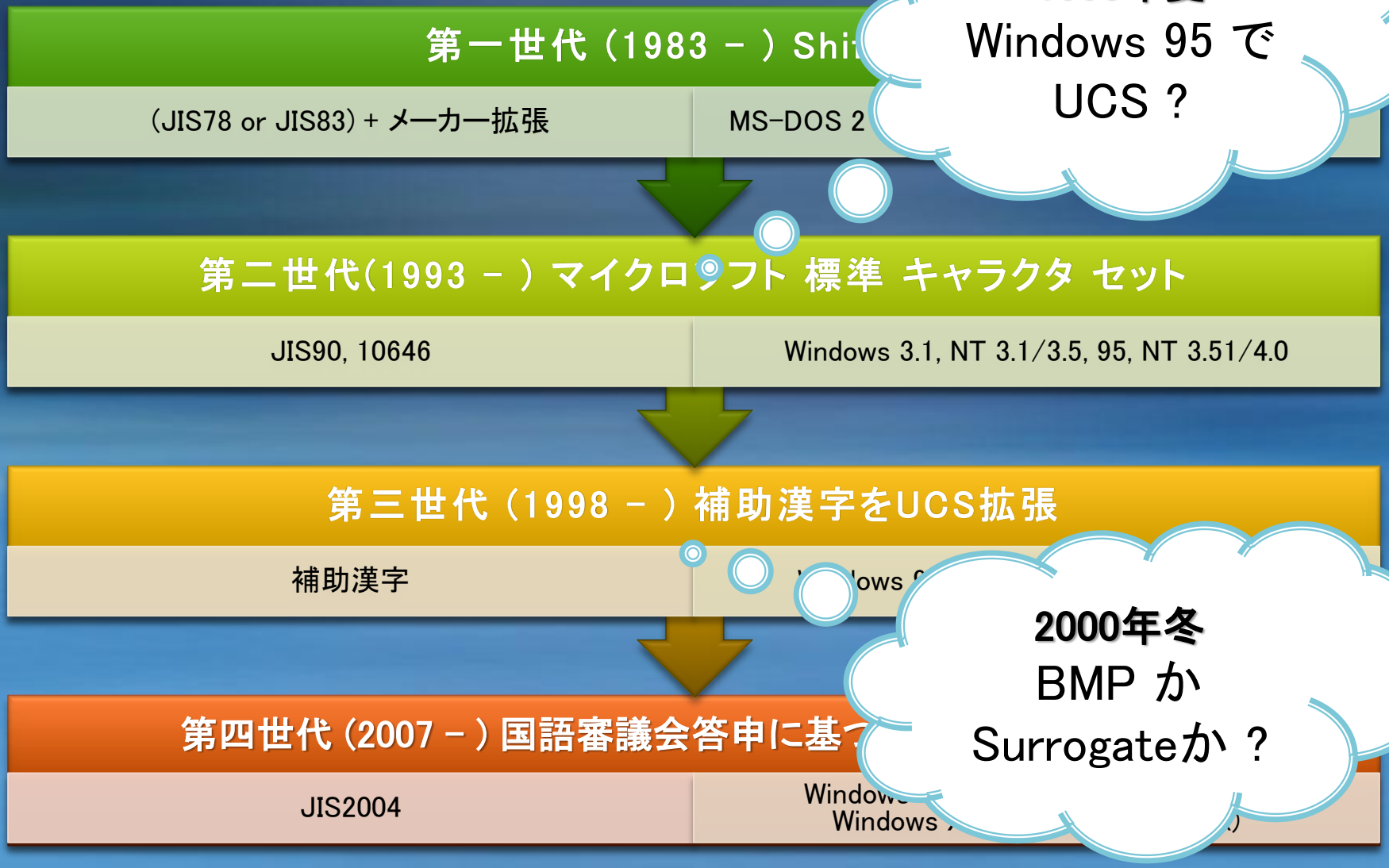
- 1997年 12月 発表
  http://www.microsoft.com/japan/presspass/detail.aspx?newsid=1260
- 1998年7月 Windows 98 で対応
- 初めてのUCSでのみの拡張
  - JIS X 0212 – 補助漢字
  - JIS X 0221 の日本語レパートリーの記号
- 関連した機能追加
  - UI や ファイル名で使われていた半角カタカナの撲滅
  - UI 向けのフォント（MS UI ゴシック）

第四世代

国語審議会答申に基づく最新規格対応 ~ JIS2004 – Surrogate Pairと字形変更 ~

- 2005年 7月 発表
  http://www.microsoft.com/japan/presspass/detail.aspx?newsid=2353
- 2007年1月 Windows Vista で対応開始
- JIS2004に対応
  - 国語審議会の答申に基づく例示字体の変更と追加漢字を反映
  - 追加漢字はSurrogate Pairも含む
- 移行のための互換性パック
  - Windows XP 向けのJIS2004対応
  - Windows Vista 向けのJIS90対応
- 関連した機能追加
  - 日本語 ClearTypeフォント「メイリオ」のUIへの採用、しかし システムデフォルトとしての対応は行わず

二つの議論

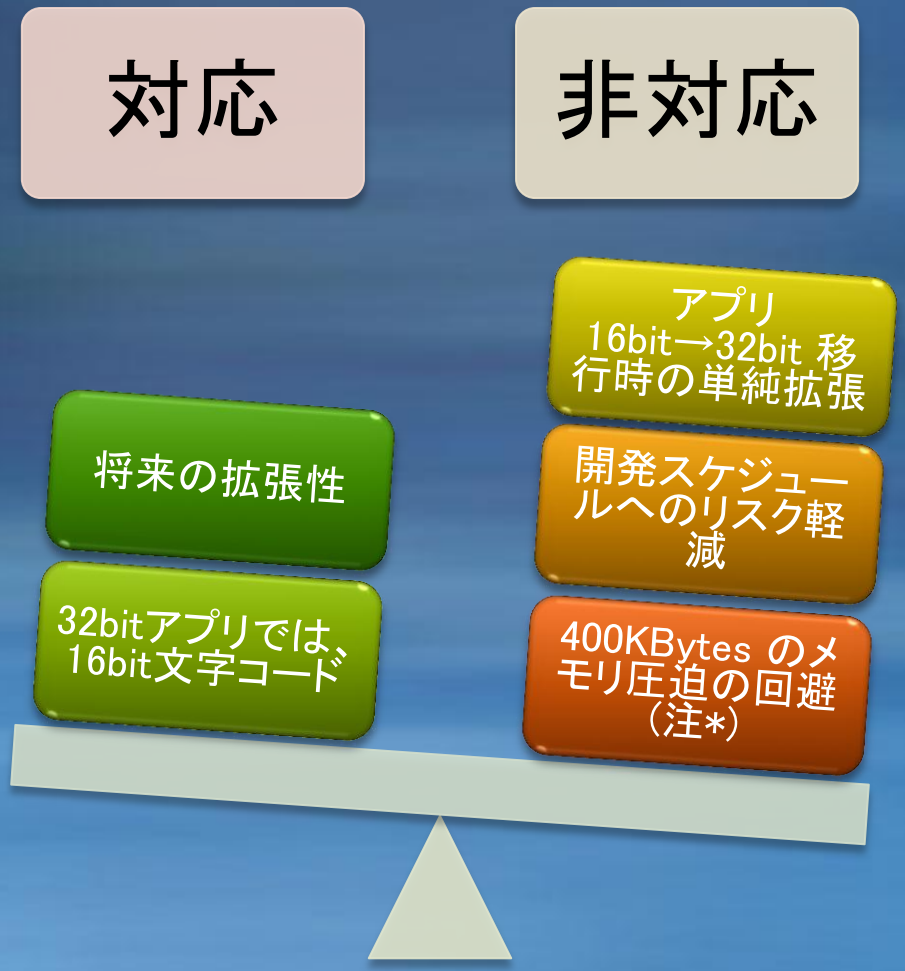


1993年夏
Windows 95 で
UCS ？

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

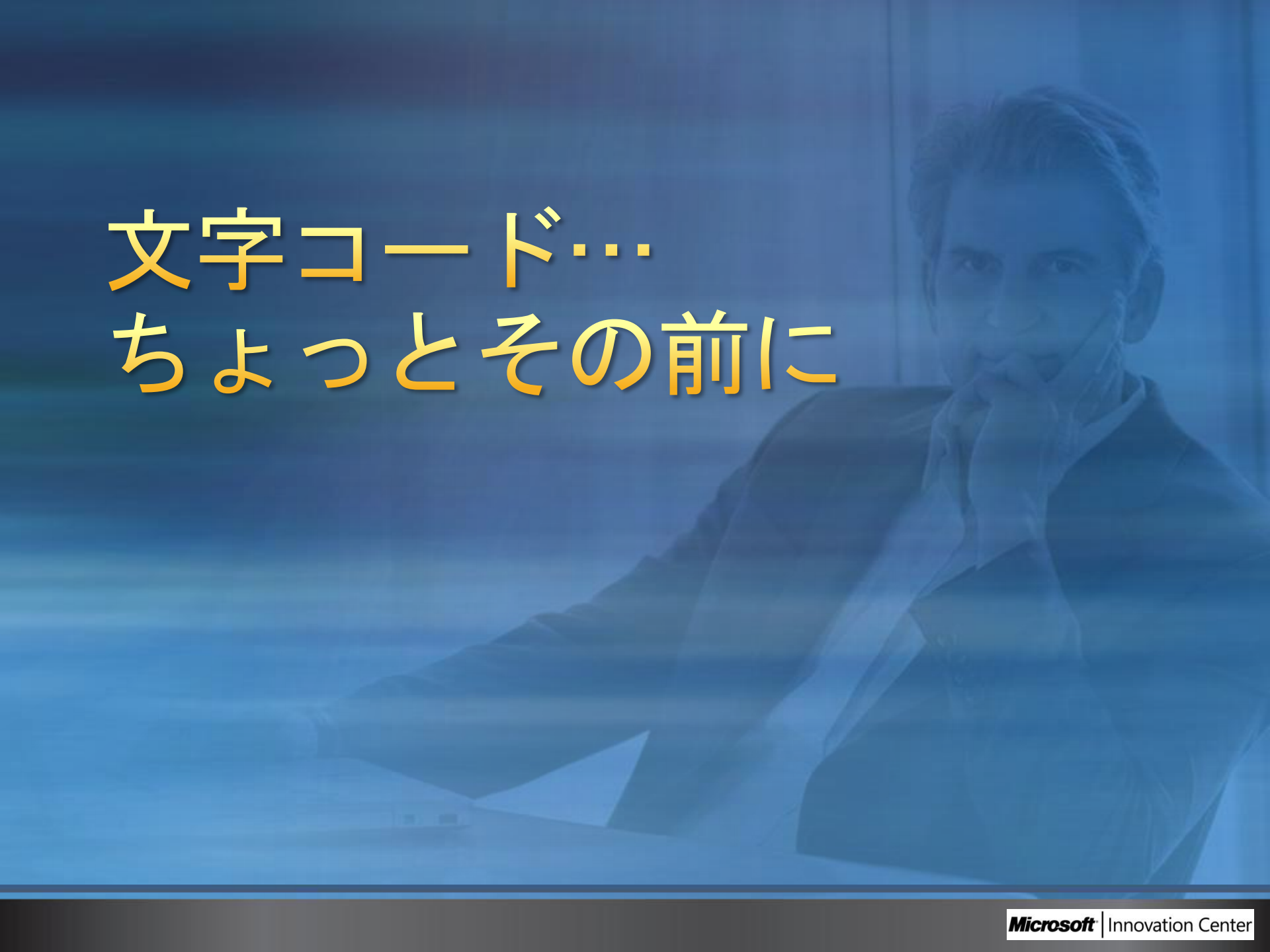


注*: 2 ~ 4Mbytes の必要メモリを想定

JIS2000 の 303文字の行方

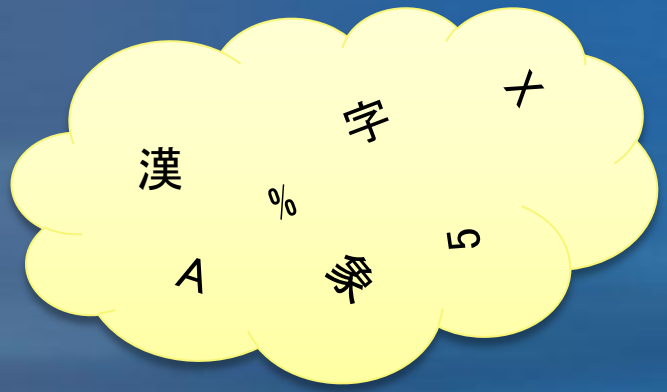
2000年 冬@ 米国 カリフォルニア州、 中国 北京

文字コード…
ちょっとその前に

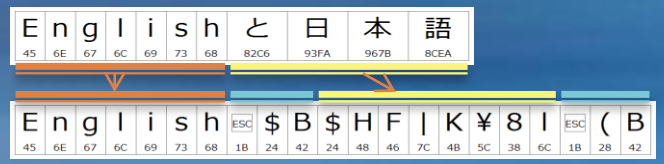


符号化文字集合、符号化方式

- 符号化文字集合
  - 文字の集合を定義



- 文字符号化方式
  - 文字集合をデータ交換可能な共通に定義された方法によりコード化
  - Shift-JIS, UTF-8, UTF-16など



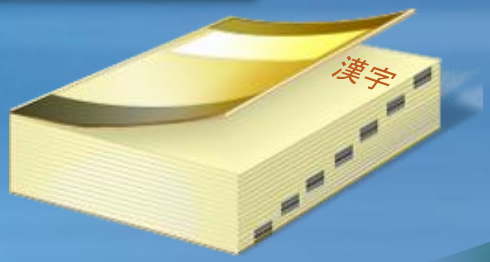
「	丈	丄	堅	」	は	追	加	面	の	文	字
300C	D840 DC0B	D844 DE3D	D844 DF1B	300D	306F	8FFD	52A0	9762	306E	6587	5B57

符号化文字集合 v.s. 符号化方式

文字とコードが一体で
定義、使用されていた

文字と符号化方式が
別々に議論されるように

符号化文字集合と
符号化方式



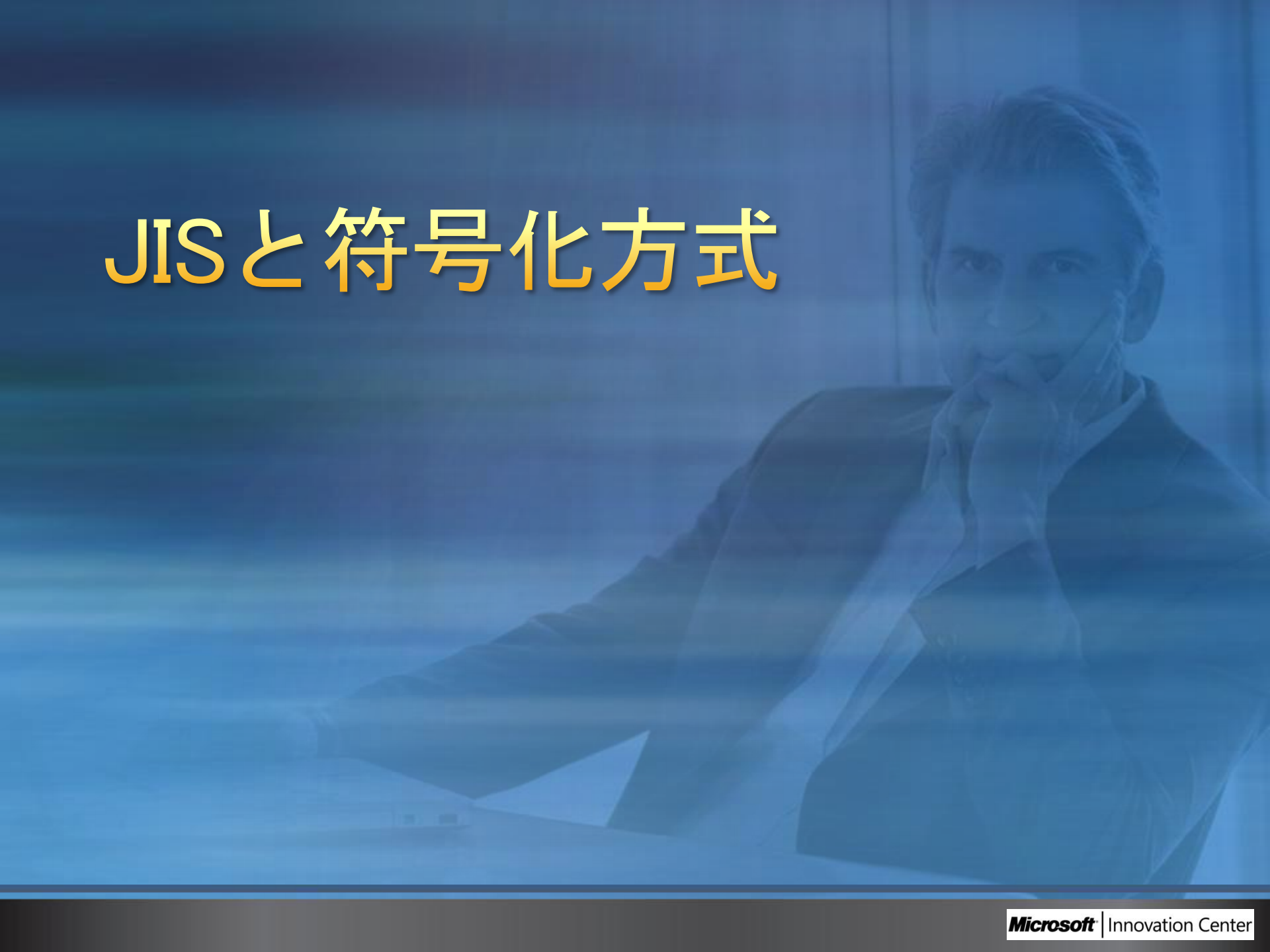
符号化文字集合

Shift-JIS

iso-2022-jp

UTF-8

JISと符号化方式



JIS X の歴史

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

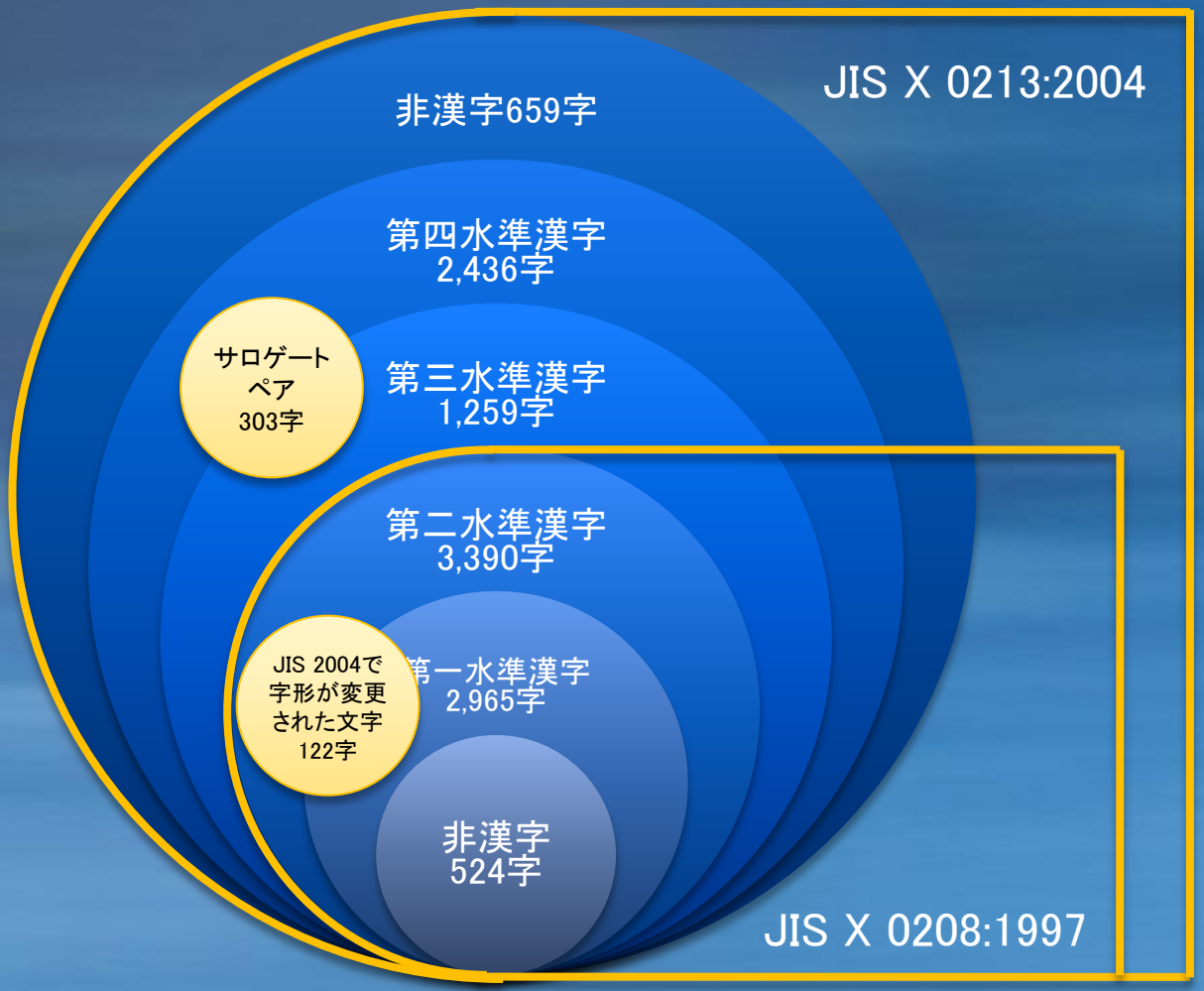
6,879字

2004 印刷標準字体

11,233字

Unicodeの対応

サロゲートペアの使用



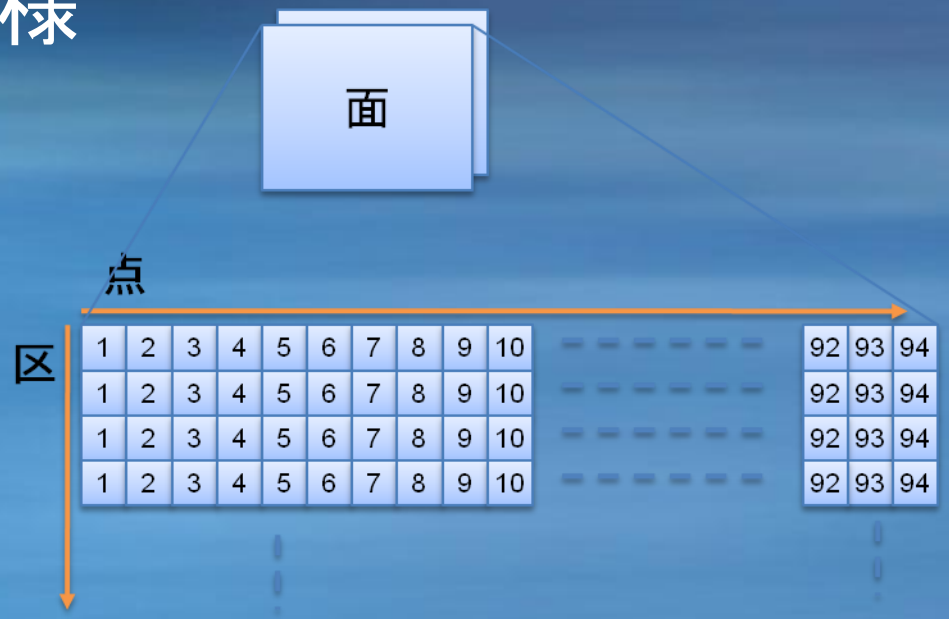
JIS2004における大きな変更

- 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に搭載されているOpenTypeフォントは、両方の字体をサポートしています。

文字集合の管理

- 区、点、そして面にて文字を管理
  - 94(区) x 94(点) x n(面)[x, y, z]
  - 8,836文字 / 面
- 符号化はこの区点面[x, y, z]の変換に関する仕様



Unicodeと符号化方式

そもそもUnicodeとは

About the Unicode Standard

The Unicode Standard is a character coding system designed to support the worldwide interchange, processing, and display of the written texts of the diverse languages and technical disciplines of the modern world. In addition, it supports classical and historical texts of many written languages.

言語、国を超えた相互運用性、相互接続性の実現を文字コード、表示そしてそのシステムデザインの見地からアプローチ

Unicodeのこれまで

Unicode バージョン	制定年	詳細
1.1	1992	• JIS X 0208 と JIS X 0212 を含む Unicode のバージョン
2.0	1996	• 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	• ユーロ通貨記号追加、多少数の記号定義変更
3.0	1999	• CJK 統合漢字拡張 A、漢字 6,582 文字追加
3.1	2001	• 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	• 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 (～ ⑱) などの非漢字を追加
4.0.0	2003	• 異体字セレクタ 17 ～ 256 追加
5.0.0	2006	• 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

符号化方式

名称	最大値	符号長（バイト）
UTF-8	U+10FFFF	1 ～ 4
UTF-16	U+10FFFF	1 ～ 2
UTF-16BE	U+10FFFF	1 ～ 2
UTF-16LE	U+10FFFF	1 ～ 2
UTF-32	U+10FFFF	4

2バイトの最大値は65,535だが・・・

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

U+10FFFF

U+EFFFD

Plane 14 (Tertiary 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

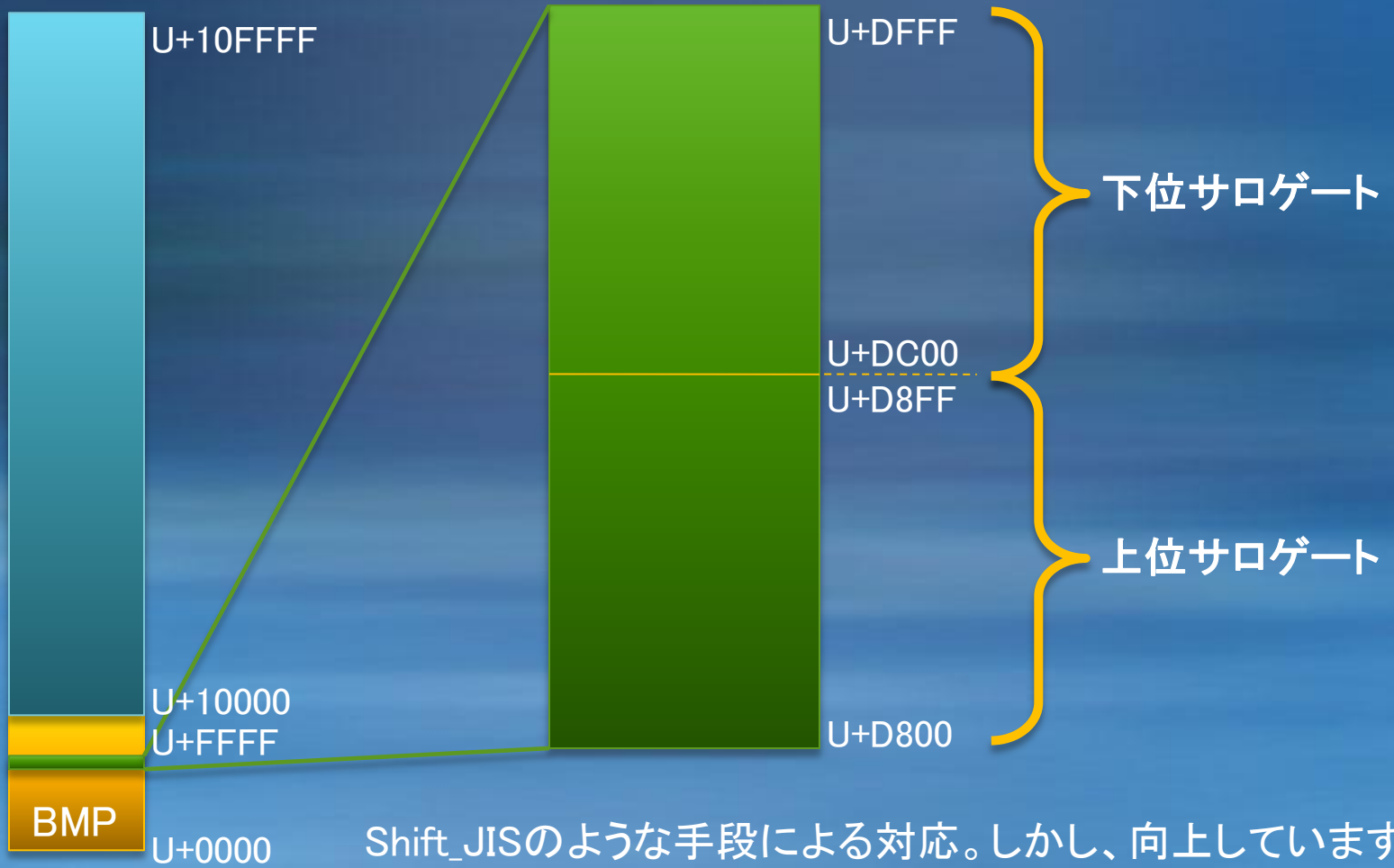
U+FFFF

Plane 0 (Basic Multilingual Plane)

U+0000

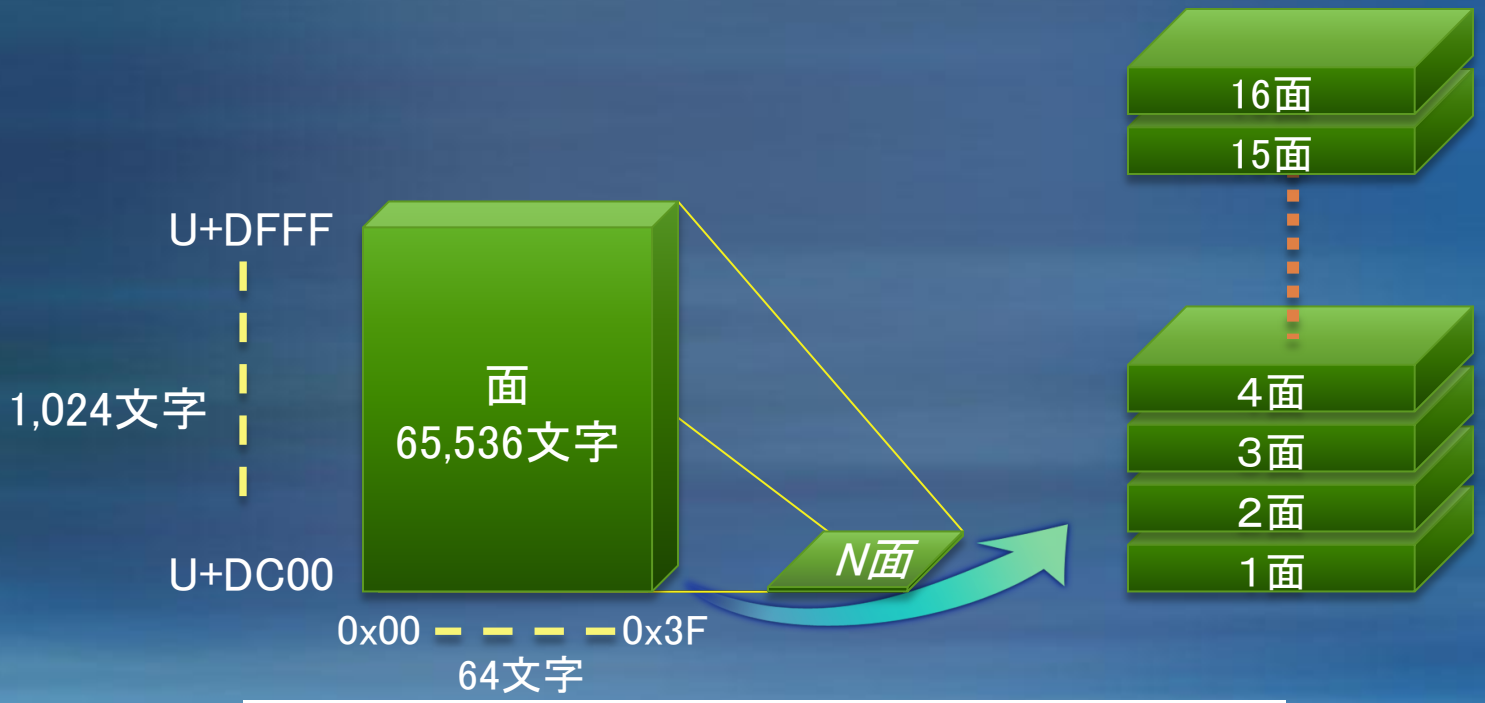
http://unicode.org/roadmaps/index.html

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	丈	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)		合計拡張数
1,024	X	1,024	=	1,048,576文字

Unicodeが定義する文字とは

- 基底文字、結合文字、合成済み文字、そして合成列・・・
  - 見た目が同じ文字、または文字の組み合わせ
  - 見た目は異なるが意味的にはほぼ同じ文字、または文字の組み合わせ
  - 見た目は同じだが定義が異なる文字

見た目が同じ

- 文字の見た目、意味共に全く同じ
  - しかしコード上は異なる

が U+304C

が = か U+304B + ゙ U+3099

- 文字列を比較した時の期待される結果は？
- どのようなAPIを使用するべき？

ë U+0451

ë = e U+0435 + ¨ U+0308

正規等価

見た目が同じだが・・・

- 見た目は同じに見えますが・・・
- 比較した場合の結果は？
- どのように扱うべきか？

U+2460 CIRCLED DIGIT ONE ≈ <circle> 0031 1

U+2780 DINGBAT CIRCLED SANS-SERIF DIGIT ONE

U+0041 LATIN CAPITAL LETTER A

U+0391 GREEK CAPITAL LETTER ALPHA

意味がほぼ同じ

- 見た目は若干異なる
- 意味的にはほぼ同じ

互換等価

多言語対応

- 初期の文字コード
  - Printable String
  - AI5
  - T.61
- EA言語への対応
  - iso-8859-1
- 合成済文字のサポート

Combining Diacritical Marks
U+0300 ~ など

正規化

- 合成済文字、合成列の扱い
  - 正規分解 (Canonical Decomposition)
  - 正規分解、正規合成 (Canonical Decomposition, followed by Canonical Composition)
  - 互換分解 (Compatibility Decomposition)
  - 互換分解、正規合成 (Compatibility Decomposition, followed by Canonical Composition)

U+01FB LATIN SMALL LETTER A WITH RINKG ABVE AND ACUTE

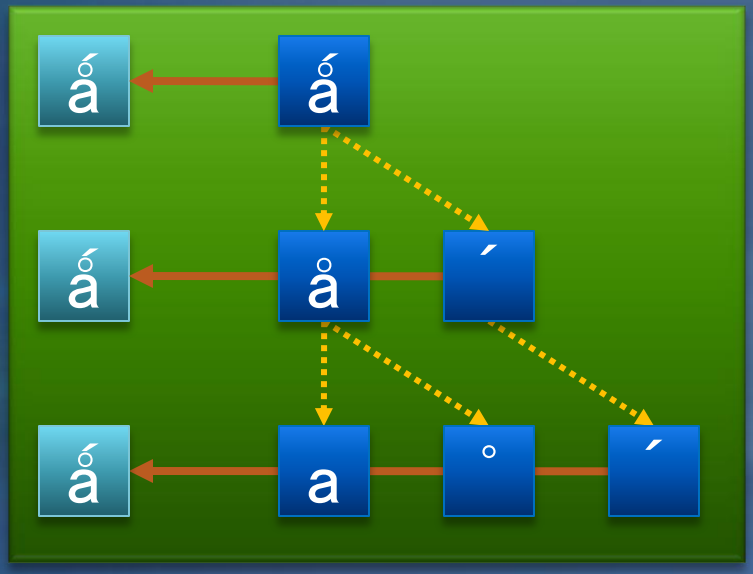
≡ 00E5  0301 

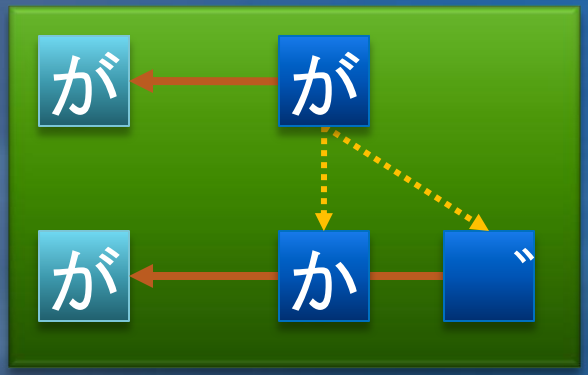
U+00E5 LATIN SMALL LETTER A WITH RING ABOVE

≡ 0061  030A 

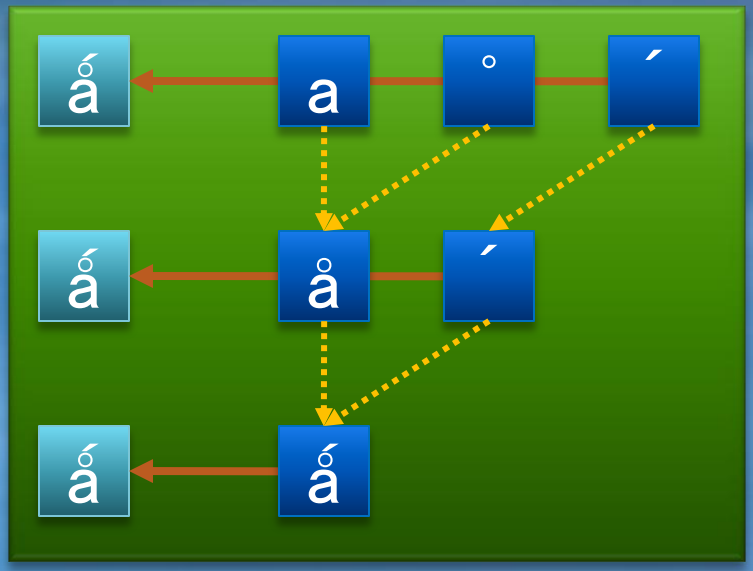
合成、分解

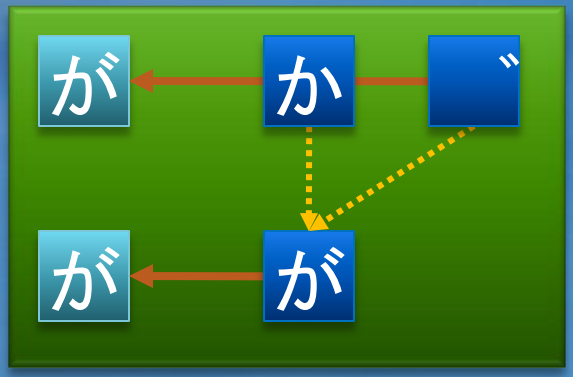
分解





合成





処理上の留意点

- 悩ましい文字データの処理
  - 保存、送信
  - 比較、並び替え

バイナリ	バ	U+30D0	イ	U+30A4	ナ	U+30CA	リ	U+30EA
ハ゛イナリ	ハ	U+30CF	゛	U+309B	イ	U+30A4	ナ	U+30CA	リ	U+30EA
バイナリ	ハ	U+30CF	゙	U+3099	イ	U+30A4	ナ	U+30CA	リ	U+30EA
ﾊﾞｲﾅﾘ	ﾊ	U+FF8A	ﾞ	U+FF9E	ｲ	U+FF72	ﾅ	U+FF85	ﾘ	U+FF98

データとして何を残す、伝える？

- 意味とオリジナルデータ
  - バイナリ上の整合性
  - 意味上の整合性

- 想定が難しいソース側（作成側、送付元）の実装とその意図
  - 「表示」、「比較」に集約される実装上の課題

- 必要な場合にのみ文字列を変換
- 可能な限りオリジナルデータを保存、送信

フォント

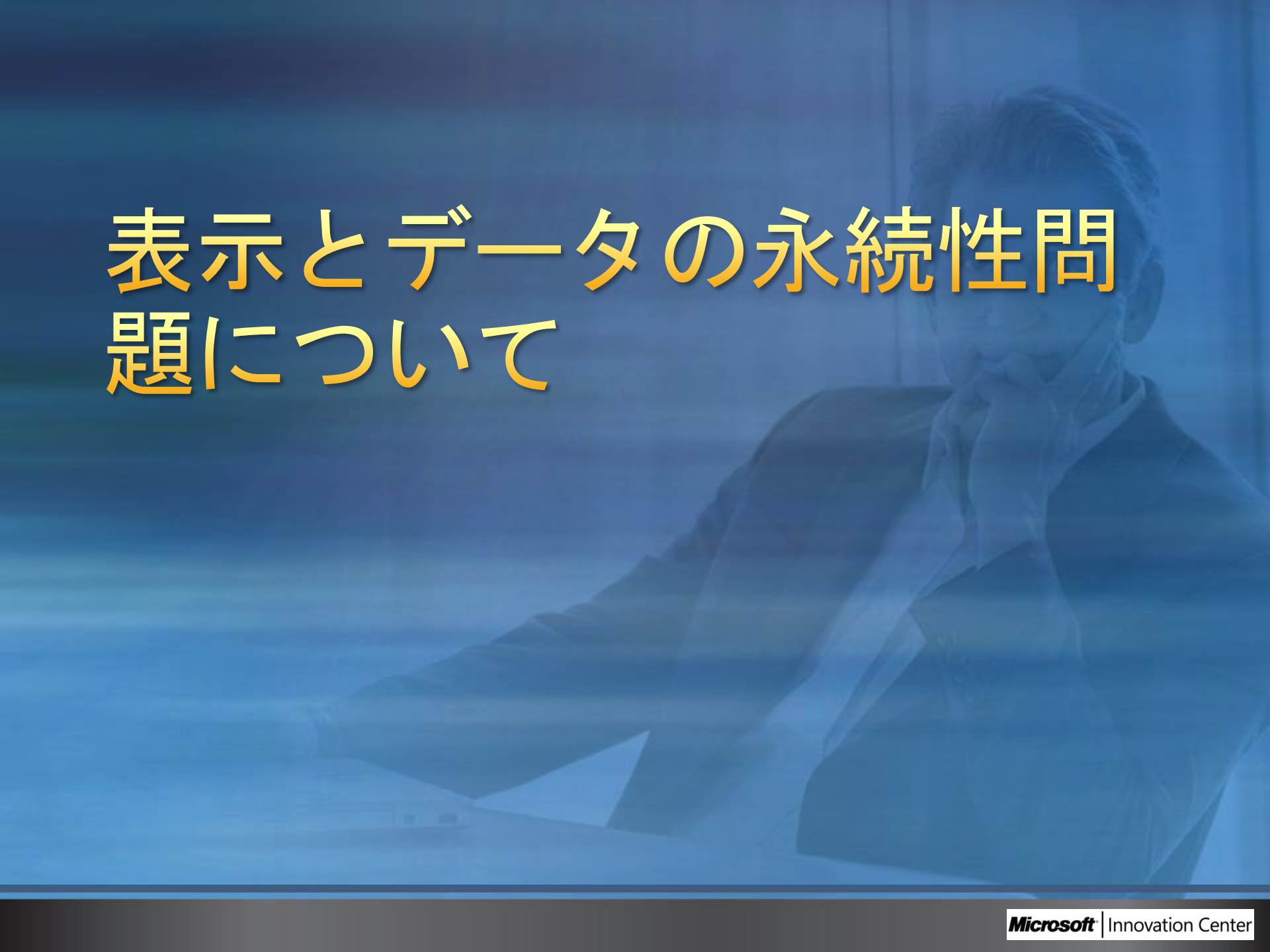
- PostScriptフォント
  - OCF (Original Composite Font)
  - CID (Character Identifier Font)
- OpenType フォント
  - PostScript
  - TrueType

※ ISO/IEC 14496-22:2009 (Second Edition)
"Open Font Format" standard

様々な字体、文字

- 地域、文化による様々な文字、様々な字体
- アドビによる取り組み
  - Adobe-Japan1-x

表示とデータの永続性問題について



システムによって異なりうる表示

- デジタル化によってバイナリデータは保存、送信可能
- 作成者の意図と表示
  - オリジナルデータと作成者の意図
  - 異なるシステム、アプリケーションによる表示

U+5473, U+5641⇒ 味噌 ？それとも 味噌 ？

- 時間、OSを越えて同じ文字を表示出来ない
  - JIS90 & JIS 2004
  - JIS20xx…？

様々な字体への対応

- ユーザー定義文字 (User Defined Character)
  - JIS X 0201、JIS X 0208に含まれない文字への使用が一般化
  - ベンダーによって異なる一貫性の無いコードの割り当てによる相互運用性の欠如
  - 特定のシステムでのみ有効な文字列、テキストデータ

IVS (Ideographic Variation Sequence)
IVD (Ideographic Variation Database)

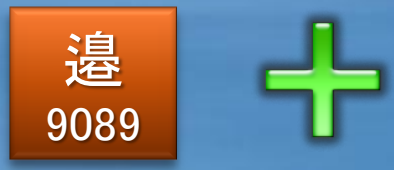
異体字セレクタ

- IVD (Ideographic Variation Database)
- IVS (Ideographic Variation Sequence)
- 「傑作」
  - U+5091, U+E0100, U+4F5C

http://unicode.org/reports/tr37/

異体字セレクタの仕組み

- 基本となる文字と、その字体を指定する事が可能に
- 見た目の事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



基本となる文字　　U+9089のバリエーション

16ビットから64ビット
(UTF-16において)

- 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

将来に向けて

- 普遍のテーマ
  - 安定した符号化文字集合は、ICT機器間の相互運用性確保の根幹となる基本要件の一つ
- 迫られるユーザー定義文字問題への対応
  - 早期の全廃が課題（相互運用性の妨げ）

- 国際標準と国内標準と国際競争力
  - Unicode
  - JIS X
  - ユーザー定義文字

最後に

～Microsoft Innovation Centerについて～

- ソフトウェア開発、検証支援
- 技術、ビジネスセミナー

http://www.microsoft.com/japan/mic

- 350台超のサーバーマシン（96-core to 2-core / x64）
- 330台超のデスクトップマシン（x64, Quad-Core）
- 750 TB超のストレージ（SAN, iSCSI, FCoE）
- 10G Network
- WANエミュレータ、最新のネットワーク環境

- 移設前と比較し20%の電力削減を実現
- Microsoft System Centerにより消費電力、空調を各サーバーラック毎に監視、個別制御

Microsoft®

Your potential. Our passion.™