Legacy & Not-So-Legacy Character Sets & Encodings

Ken Lunde

CJKV Type Development
Adobe Systems Incorporated

Adobe

ftp://ftp.oreilly.com/pub/examples/nutshell/cjkv/unicode/iuc15-tb1-slides.pdf

Tutorial Overview

- What is a character set? What is an encoding?
- How are character sets and encodings different?
- Legacy character sets.
- Non-legacy character sets.
- Legacy encodings.
- How does Unicode fit it?
- Code conversion issues.
- *Disclaimer:* The focus of this tutorial is primarily on Asian (CJKV) issues, which tend to be complex from a character set and encoding standpoint.

Terminology & Abbreviations

- **GB (China)**
  - Stands for “Guo Biao” (国标 *guóbiāo*).
  - Short for “Guojia Biaozhun” (国家标准 *guójiā biāozhǔn*).
  - Means “National Standard.”
- **GB/T (China)**
  - “T” stands for “Tui” (推 *tuī*).
  - Short for “Tuijian” (推荐 *tuījiàn*).
  - “T” means “Recommended.”
- **CNS (Taiwan)**
  - 中國國家標準 (*zhōngguó guójiā biāozhǔn*) in Chinese.
  - Abbreviation for “Chinese National Standard.”

Terminology & Abbreviations (Cont'd)

- **GCCS (Hong Kong)**
  - Abbreviation for "Government Chinese Character Set."
- **JIS (Japan)**
  - 日本工業規格 (*nihon kōgyō kikaku*) in Japanese.
  - Abbreviation for "Japanese Industrial Standard."
  - ㉾
- **KS (Korea)**
  - 한국 공업 규격 (韓國工業規格 *hangug gongeob gyugyeog*) in Korean.
  - Abbreviation for "Korean Standard."
  - ㉿
  - Designation change from "C" to "X" on August 20, 1997.

Terminology & Abbreviations (Cont'd)

- TCVN (Vietnam)
  - — *Tiêu Chuẩn Việt Nam* in Vietnamese.
  - — Means "Vietnamese Standard."
- CJKV
  - — Chinese, Japanese, Korean, and Vietnamese.

What Is A Character Set?

- A collection of characters that are intended to be used together to create meaningful text.
- There are two varieties of character sets:
  - Small character sets (generally 256 characters or less).
  - Large character sets (generally have thousands of characters).
- All character sets have errors—nothing is perfect.
- No character set can possibly include all characters.
- Most CJKV character sets have a common structure, such as being based on a 94×94 matrix.

What Is An Encoding?

Adobe

- The systematic method of defining the correspondence between numerical codes and the final printable glyphs.
- Different encodings can represent the same character:
  - *Example:* 中 (“middle” or “center”).
  - 0x4E2D in UCS-2 or UTF-16, 0xE4B8AD in UTF-8.
  - 0x4366 in ISO-2022-JP, 0xC3E6 in EUC-JP, 0x9286 in Shift-JIS.
  - 0x5650 in ISO-2022-CN, 0xD6D0 in EUC-CN.
  - 0x4463 in ISO-2022-CN, 0xC4E3 in EUC-TW, 0xA4A4 in Big Five.
  - 0x7169 in ISO-2022-KR, 0xF1E9 in EUC-KR.

Character Sets Versus Encodings?

- Are character sets and encodings the same? Absolutely not!
- It is important to distinguish character sets and encodings.
- The "CJKV World" makes it easy to illustrate the distinction.
- Character sets can be encoded in many ways:
  - *Example:* JIS X 0208:1997 can be encoded according to ISO-2022-JP, EUC-JP, and Shift-JIS.
- Encodings can encode multiple character sets:
  - *Example:* EUC-JP encoding can encode JIS X 0201-1997, JIS X 0208:1997, and JIS X 0212-1990.

What Is A Legacy Character Set?

- Any character set that has complete support (that is, full coverage) in a newer character set.
- Context is critical—Unicode is considered the "newer" character set in the context of this talk.
- The number of legacy character sets increases over time, as the newer character set becomes larger through expanded coverage.
  - Unicode Version 2.1 → Version 3.0 adds 6,582 Chinese characters.

Legacy Character Sets

- *Definition:* Full support in Unicode.
- More and more character sets become legacy ones over time, as Unicode expands.

Legacy Character Sets (Cont'd)

- **ASCII (ANSI X3.4-1986)**
  - ISO 646:1991
- **ISO 8859 Series**
  - ASCII as the base.
  - Over 10 parts.
  - Part 1, ISO 8859-1:1998, is the most widely used.
- **CJKV flavors of ASCII (differences illustrated later):**
  - GB 1988-89 (China)
  - CNS 5205-1989 (Taiwan)
  - JIS X 0201-1997 (Japan)
  - KS X 1003:1993 (South Korea)
  - TCVN 5712:1993 (Vietnam)

Legacy Character Sets (Cont'd)

- China
- GB 1988-89
  - Equivalent to ASCII.
  - 0x24 is "yuan" (¥) instead of "dollar" ($).
- GB 2312-80
  - 7,445 characters.
  - 6,763 of which are hanzi, separated into two levels.
- GB/T 12345-90
  - Traditional analog of GB 2312-80.
- GBK
  - Includes all Unicode Version 2.1 hanzi.

Legacy Character Sets (Cont'd)

- **Taiwan**
- **CNS 5205-1989**
  - Equivalent to ASCII.
- **Big Five**
  - 13,494 characters.
  - 13,053 of which are hanzi, separated into two levels.
- **CNS 11643-1992 Planes 1 and 2**
  - Equivalent to Big Five.
  - An additional five planes are available.
  - 48,711 characters.
  - 48,027 of which are hanzi, separated into seven levels.

Legacy Character Sets (Cont'd)

- Japan
- JIS X 0201-1997
  - Equivalent to ASCII.
  - 0x5C is "yen" (¥) instead of "backslash" (\).
- JIS X 0208:1997
  - 6,879 characters.
  - 6,355 of which are kanji, separated into two levels.
- JIS X 0212-1990
  - 6,067 characters.
  - 5,801 of which are kanji.
  - May be supplanted by JIS X 0213:1999.

Legacy Character Sets (Cont'd)

- Japan
- IBM Selected Kanji
  - 28 non-kanji.
  - 360 kanji.
  - 279 of the kanji are in JIS X 0212-1990.

Legacy Character Sets (Cont'd)

- South Korea
- KS standard re-designation took place in1997:
  - KS C 5601-1992 → KS X 1001:1992
  - KS C 5657-1991 → KS X 1002:1991
  - KS C 5636-1993 → KS X 1003:1993
- KS X 1003:1993
  - Equivalent to ASCII.
  - 0x5C is "won" (₩) instead of "backslash" (\).
- KS X 1001:1992
  - 8,224 characters.
  - 2,350 of which are hangul, and 4,888 of which are hanja.

Legacy Character Sets (Cont'd)

- Vietnam
- TCVN 5712:1993
  - — ASCII plus 139 characters for Quốc ngữ.
- TCVN 6056:1995
  - — 3,311 characters, all of which are chữ Hán.

Common Character Sets

- Ignoring “small” character sets, such as ASCII and its equivalents.
- China
  - GB 2312-80
  - GBK (GB 2312-80 is a pure subset, from both a character set and encoding perspective)
- Taiwan
  - Big Five
  - CNS 11643-1992
- Japan
  - JIS X 0208:1997

Common Character Sets (Cont'd)

- South Korea
  - KS X 1001:1992

GB 2312-80 Issues

Adobe

- **Extensions, specified in other standards:**
  - GB 6345.1-86 added 6 pinyin, half-width GB 1988-89, and 32 half-width pinyin.
  - GB 8565.2-88 added 636 hanzi plus 69 non-hanzi.
  - ISO-IR-165:1992 is identical to GB 6345.1-86 and GB 8565.2-88, plus 138 more hanzi.
- **Corrections, specified in another standard:**
  - In GB 6345.1-86
  - 03-71 correction (g should be g)
  - 79-81 correction (鍾 should be 锺)

GB 2312-80 Issues (Cont'd)

Adobe

- Corrections, unspecified:
  - The ordering of Cyrillic Ф and Х was reversed in first—and possibly second—printing of the standard.

GB/T 12345-90 Issues

Adobe

- Traditional analog of GB 2312-80.
  - 2,118 traditional form replacements for rows 16 through 87.
  - 62 additional traditional forms in rows 88 and 89.
  - See pp 897–916 of *CJKV Information Processing*.
- 103 additional hanzi in rows 88 and 89.
  - 41 are simplified forms moved from rows 16 through 87.
  - See pp 915–916 of *CJKV Information Processing*.
- Corrections:
  - 33-05 (隸 should be 隸)
  - 57-76 (鳧 should be 鳧)

GB/T 12345-90 Issues (Cont'd)

Adobe

- Conflicting/ambiguous Unicode mappings.
  - 22 such mappings.
  - See Table 3-28 on page 86 of *CJKV Information Processing*.

Big Five Issues

- **Two duplicate hanzi:**
  - — 兀 at 0xA461 and 0xC94A
  - — 殼 at 0xDCD1 and 0xDDFC
  - — Corrected in CNS 11643-1992 through removal.
- **Ordering of hanzi is different from CNS 11643-1992**
  - — Big Five Level 1 and CNS 11643-1992 Plane 1 have eight characters that are ordered differently.
  - — Big Five Level 2 and CNS 11643-1992 Plane 2 have 17 characters that are ordered differently.
  - — See Tables 3-50 and 3-51 on pp 97–98 of *CJKV Information Processing*.
  - — Consequently, code conversion between Big Five and EUC-TW (an encoding for CNS 11643-1992) is table-driven.

Big Five Issues (Cont'd)

Adobe

- Does not include the classical radicals
  - CNS 11643-1992 Plane 1 includes 213 classical radicals in rows 7 through 9.

JIS X 0208:1997 Issues

- Glyph differences between JIS C 6226-1987 (JIS78) and JIS X 0208-1983 (JIS83):
  - *Example:* 16-02 (唖 → 啞)
  - See pp 918–922 of *CJKV Information Processing*.
- Simplified/traditional character swapping between JIS78 and JIS83:
  - *Example:* 25-60 (礦 → 砿)
  - *Example:* 66-72 (砿 → 礦)
- JIS X 0208-1990 additions:
  - 84-05 (凜)
  - 84-06 (熙)

JIS X 0212-1990 Issues

- 28 duplicate kanji with JIS C 6226-1978 (aka, JIS78):
  - *Example:* 21-64 (唖) and JIS78 16-02 (唖)
  - See pp 924–925 of *CJKV Information Processing*.
- One duplicate character with JIS X 0208 (all versions):
  - 16-17 (〆) and JIS X 0208 01-26 (〆)

KS X 1001:1992 Issues

Adobe

- 268 duplicate hanja
  - Due to multiple readings for hanja.
  - *Example:* 樂 (read *nag*, *rag*, *ag*, or *yo*) has four code points (49-66, 53-05, 68-37, and 72-89).
  - See pp 933–940 in *CJKV Information Processing*.
- Includes only 2,350 hangul
  - 8,822 more are necessary to cover the possible set of 11,172 hangul.
  - Johab and UHC (Unified Hangul Code) encodings support all 11,172 hangul. Of course, Unicode Version 2.0 and later, too.

Character Sets As Glyph Standards

- Character sets typically do not serve as glyph standards—it is beyond their specification.
- However, because glyph guidance is necessary for building font products, character set standards become *de facto* glyph standards—errors are propagated.
- JIS X 0208:1997 is unique in that it illustrates possible/acceptable glyphs for many characters.
  - Example: 16-02 (唖 or 啞 are acceptable forms)
- There are GB standards that specify glyphs, based on typeface design.
  - *Example:* GB 6345.1-86

Cross-Platform Assumptions

- When working in a cross-platform environment—MacOS, Windows, and Unix—it is possible to assume a minimum level of character set support for each locale.
- ASCII everywhere
- GB 2312-80 for China and Singapore
  - Remember that GBK is not available on MacOS/Unix!
- Big Five for Taiwan and Hong Kong
  - CNS 11643-1992 is not implemented on MacOS/Windows!
- JIS X 0208:1997 for Japan
- KS X 1001:1992 for South Korea

Less Common Character Sets

- No or partial support in Unicode.
- Not-so-legacy character sets.
- China
  - GB 7589-87 (aka, GB2) with 7,237 simplified hanzi
  - GB 7590-87 (aka, GB4) with 7,039 simplified hanzi
  - GB 13131-9X (aka, GB3)—traditional analog of GB 7589-87
  - GB 13132-9X (aka, GB5)—traditional analog of GB 7590-87
  - *Note:* "Even" designations use simplified hanzi, and "odd" ones are traditional analogs.
  - All of the above have partial support in Unicode.

Less Common Character Sets (Cont'd)

- Taiwan
  - CNS 11643-1992 Planes 3 through 7 have partial support in Unicode
  - 34,976 hanzi in these five planes.
- Hong Kong
  - GCCS support in Unicode is partial, approximately 50 percent.
- Japan
  - JIS X 0213:1999 (not-yet-published)
  - *Note:* Status as legacy character set is unknown.

Less Common Character Sets (Cont'd)

- North Korea
  - KPS 9566-97
  - *Note:* Status as legacy character set is unknown—need more information.
- South Korea
  - KS X 1002:1991
  - All hanja and modern hangul are in Unicode.
- Vietnam
  - TCVN 5773:1993
  - Most chữ Nôm are not in Unicode, even Version 3.0.
- Many, many vendor extensions.

Locale-Independent Encodings

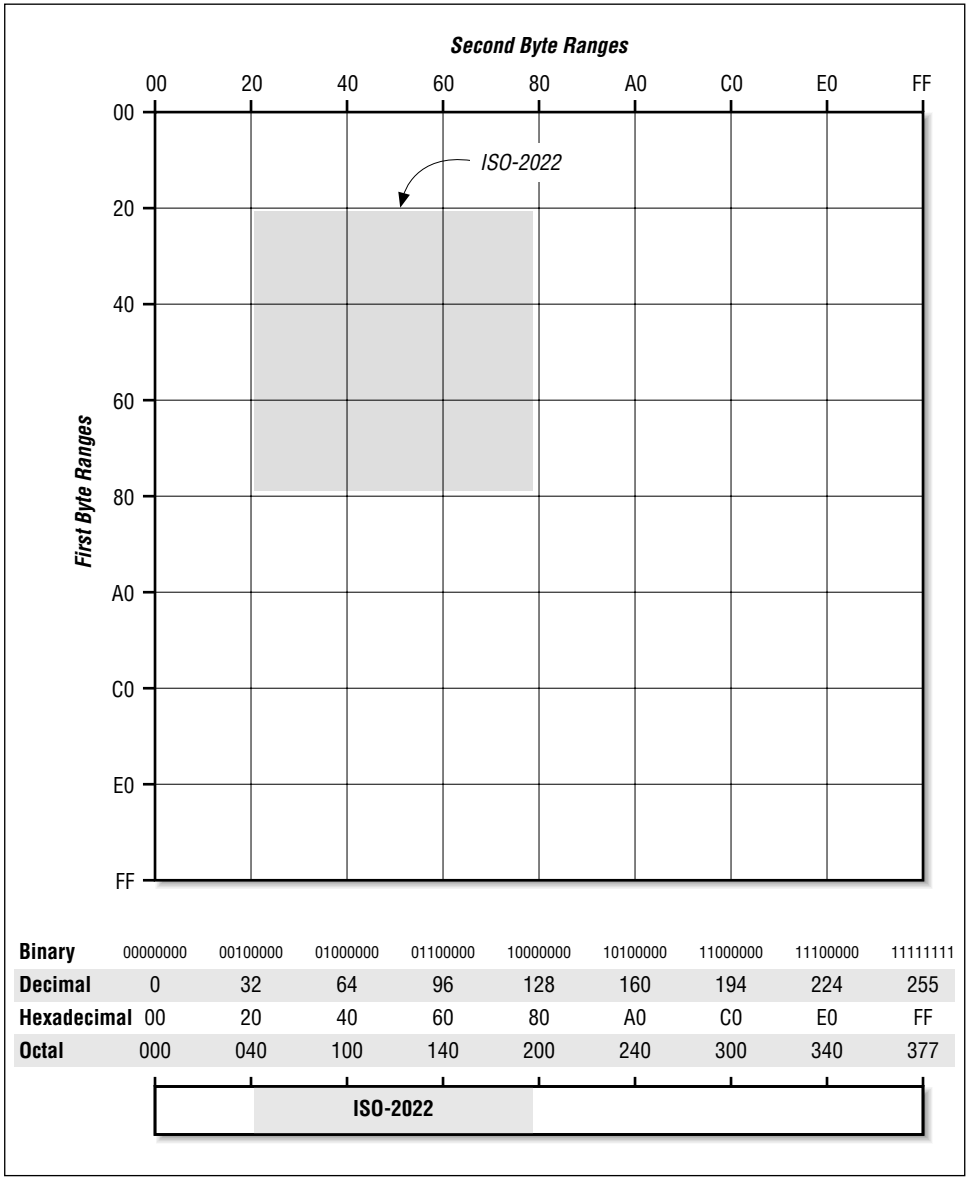
- ISO 2022 (7-bit) encodings
  - Locale-specific instances.
  - Modal encoding, using escape sequences, shifting characters, single-shift sequences, and designator sequences.
  - One- and two-byte representations.
- EUC (8-bit) encodings
  - Locale-specific instances.
  - Non-modal, variable-length encoding.
  - One- through four-byte representations.

ISO 2022 (7-bit) Encodings

- **ISO-2022-CN (China and Taiwan)**
  - — Supports ASCII, GB 2312-80, and CNS 11643-1992 Planes 1 and 2.
- **ISO-2022-CN-EXT (China and Taiwan)**
  - — ISO-2022-CN plus support for GB/T 12345-90, GB 7589-87, GB 7590-87, ISO-IR-165:1992, and CNS 11643-1992 Planes 3 through 7.
- **ISO-2022-JP (Japan)**
  - — Supports ASCII, JIS-Roman, JIS C 6226-1978, and JIS X 0208-1983.
- **ISO-2022-JP-1 (Japan)**
  - — ISO-2022-JP plus support for JIS X 0212-1990.

ISO 2022 (7-bit) Encodings (Cont'd)

- **ISO-2022-JP-2 (Japan)**
  - — ISO-2022-JP plus support for GB 2312-80, JIS X 0212-1990, KS X 1001:1992, ISO 8859-1:1998, and ISO 8859-7-1998.
- **ISO-2022-KR (South Korea)**
  - — Supports ASCII and KS X 1001:1992.
- **ISO-2022-VN (Vietnam)**
  - — Supports TCVN 5712:1993, TCVN 5773:1993, and TCVN 6056:1995.
  - — Suggested designation only—does not yet exist.



Binary	00000000	00100000	01000000	01100000	10000000	10100000	11000000	11100000	11111111
Decimal	0	32	64	96	128	160	194	224	255
Hexadecimal	00	20	40	60	80	A0	C0	E0	FF
Octal	000	040	100	140	200	240	300	340	377

EUC (8-bit) Encodings

Adobe

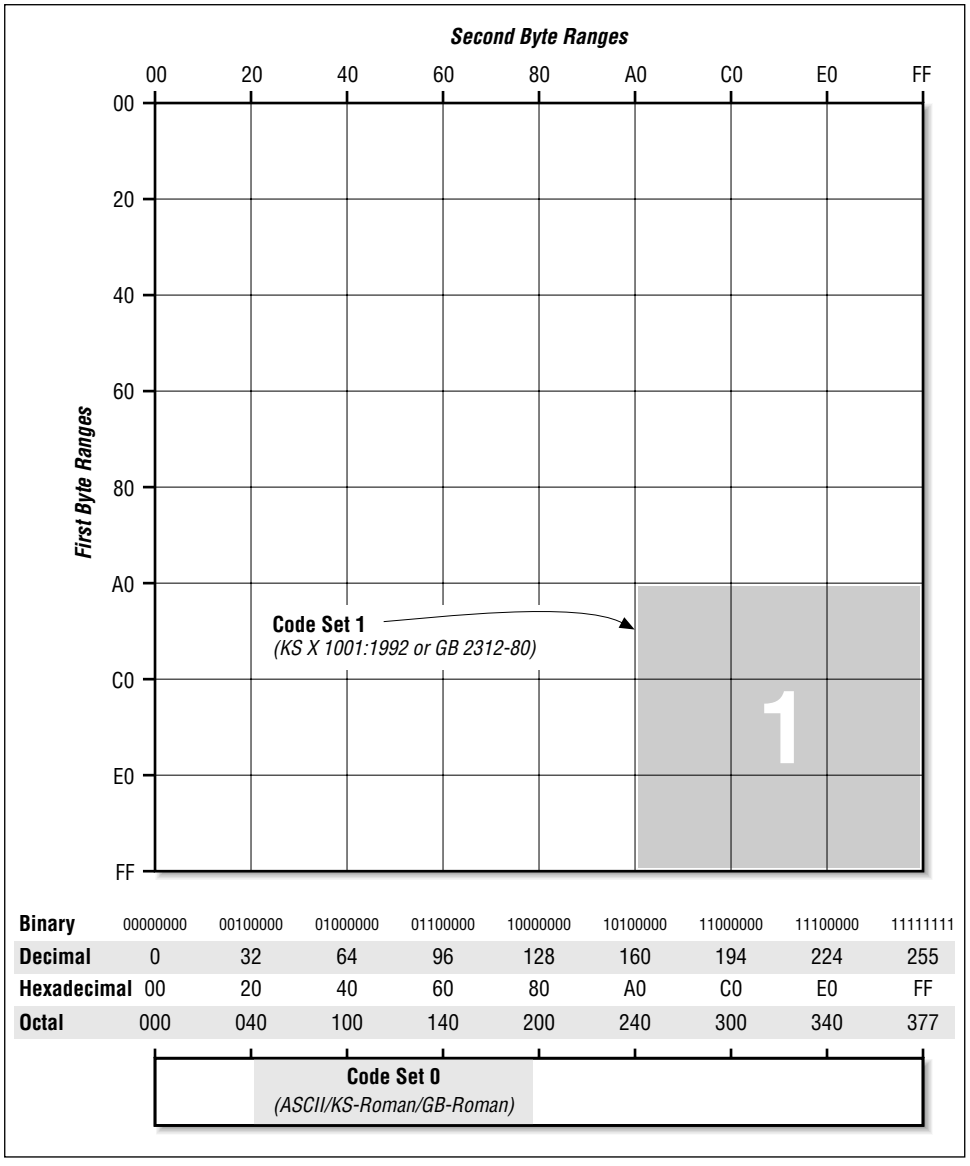
- EUC-CN (China)
  - Mixed one- and two-byte encoding.
  - GB 1988-89 or ASCII in one-byte region (0x20–0x7E).
  - GB 2312-80 in two-byte region (0xA1A1–0xFEFE).
- EUC-TW (Taiwan)
  - Mixed one-, two-, and four-byte encoding.
  - CNS 5205-1989 or ASCII in one-byte region (0x20–0x7E).
  - CNS 11643-1992 Plane 1 in two-byte region (0xA1A1–0xFEFE).
  - CNS 11643-1992 Planes 1 through 7 in four-byte region (0x8EA1A1A1–0x8EB0FEFE).
  - *Note:* CNS 11643-1992 Plane 1 duplicately encoded in two- and four-byte region.

EUC (8-bit) Encodings (Cont'd)

- EUC-JP (Japan)
  - Mixed one-, two-, and three-byte encoding.
  - JIS X 0201-1997 Latin or ASCII in one-byte region (0x20–0x7E).
  - JIS X 0208:1997 in two-byte region (0xA1A1–0xFEFE).
  - JIS X 0201-1997 half-width katakana in two-byte region (0x8EA1–0x8EDF).
  - JIS X 0212-1990 in three-byte region (0x8FA1A1–0x8FFEFE).
- EUC-KR (South Korea)
  - Mixed one- and two-byte encoding.
  - KS X 1003:1993 or ASCII in one-byte region (0x20–0x7E).
  - KS X 1001:1992 in two-byte region (0xA1A1–0xFEFE).

EUC (8-bit) Encodings (Cont'd)

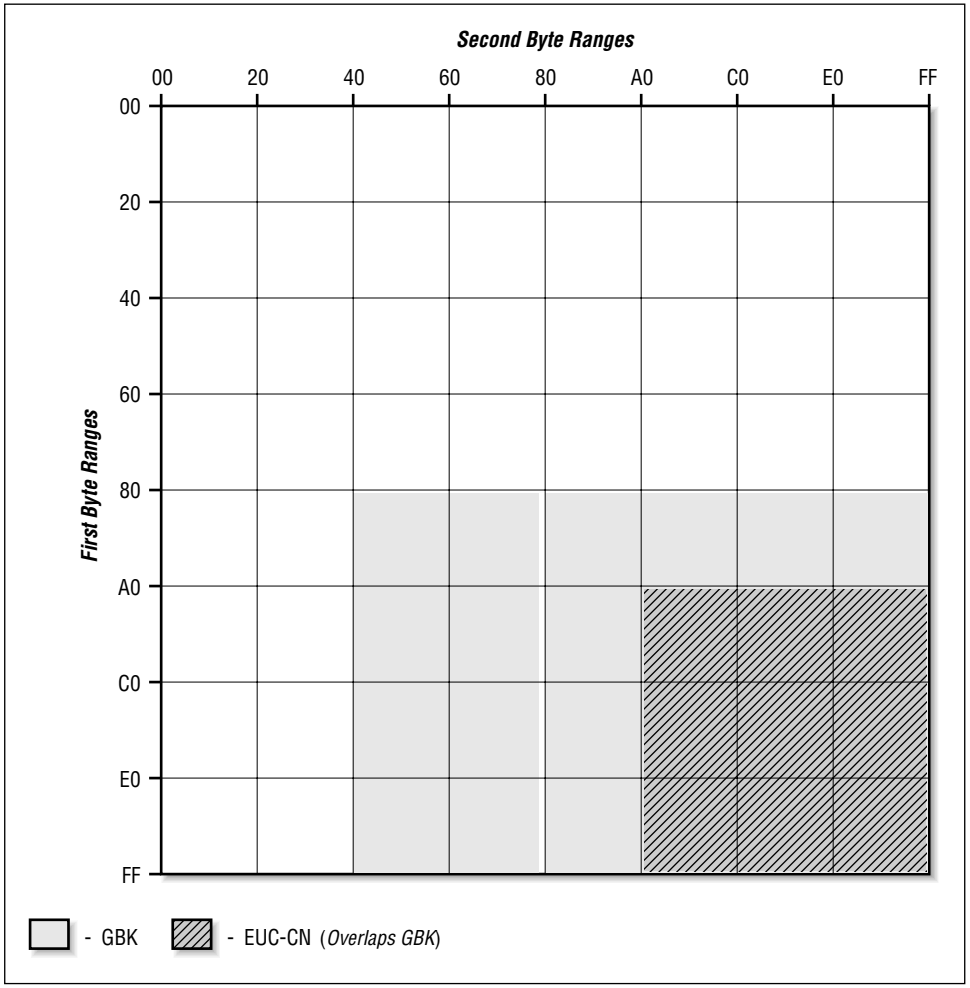
- EUC-VN (Vietnam)
  - — Supports TCVN 5712:1993, TCVN 5773:1993, and TCVN 6056:1995.
  - — Suggested designation only—does not yet exist.

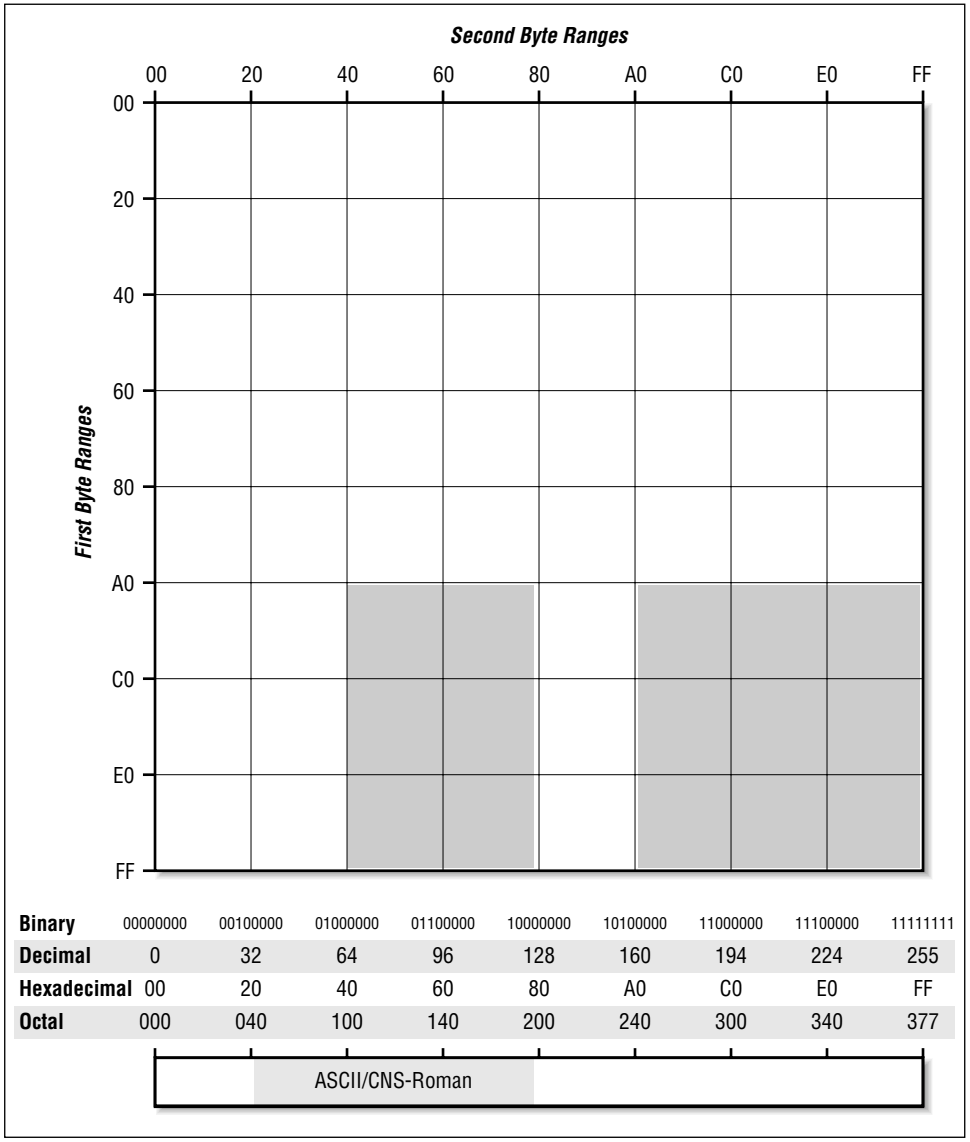


Binary	00000000	00100000	01000000	01100000	10000000	10100000	11000000	11100000	11111111
Decimal	0	32	64	96	128	160	194	224	255
Hexadecimal	00	20	40	60	80	A0	C0	E0	FF
Octal	000	040	100	140	200	240	300	340	377

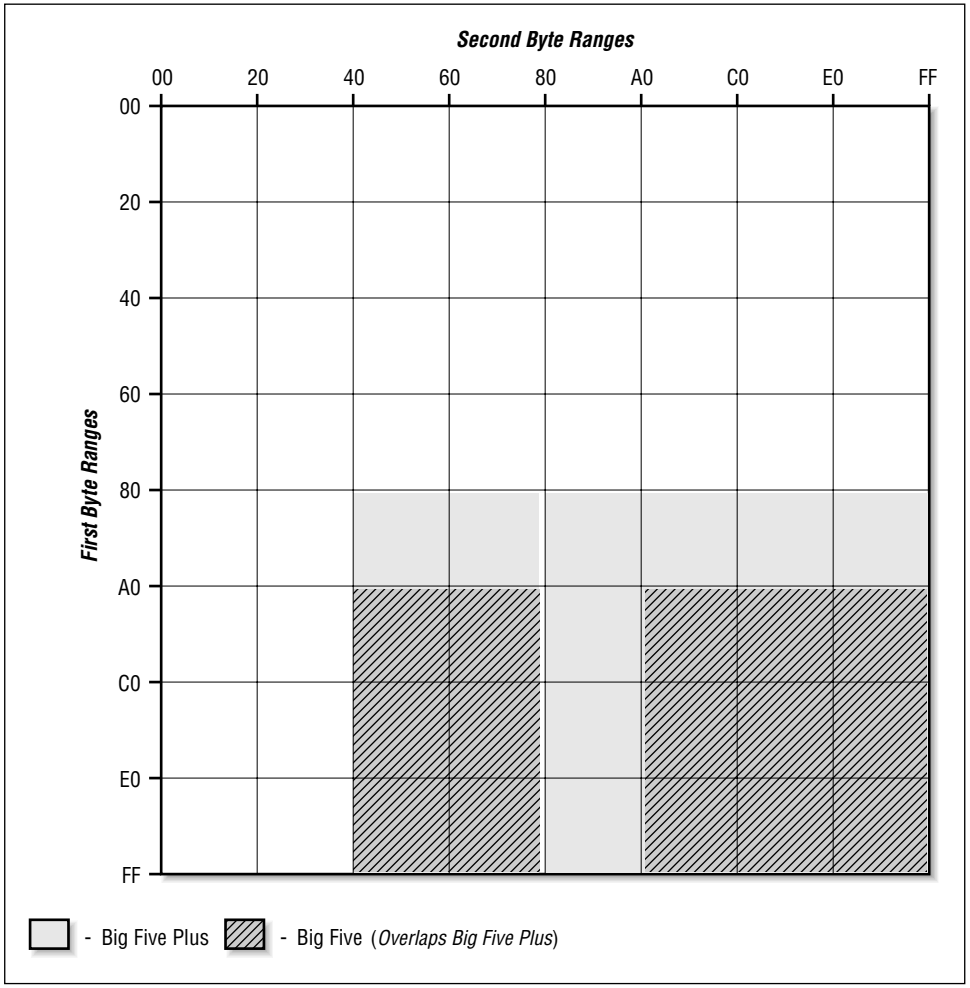
Locale-Specific Encodings

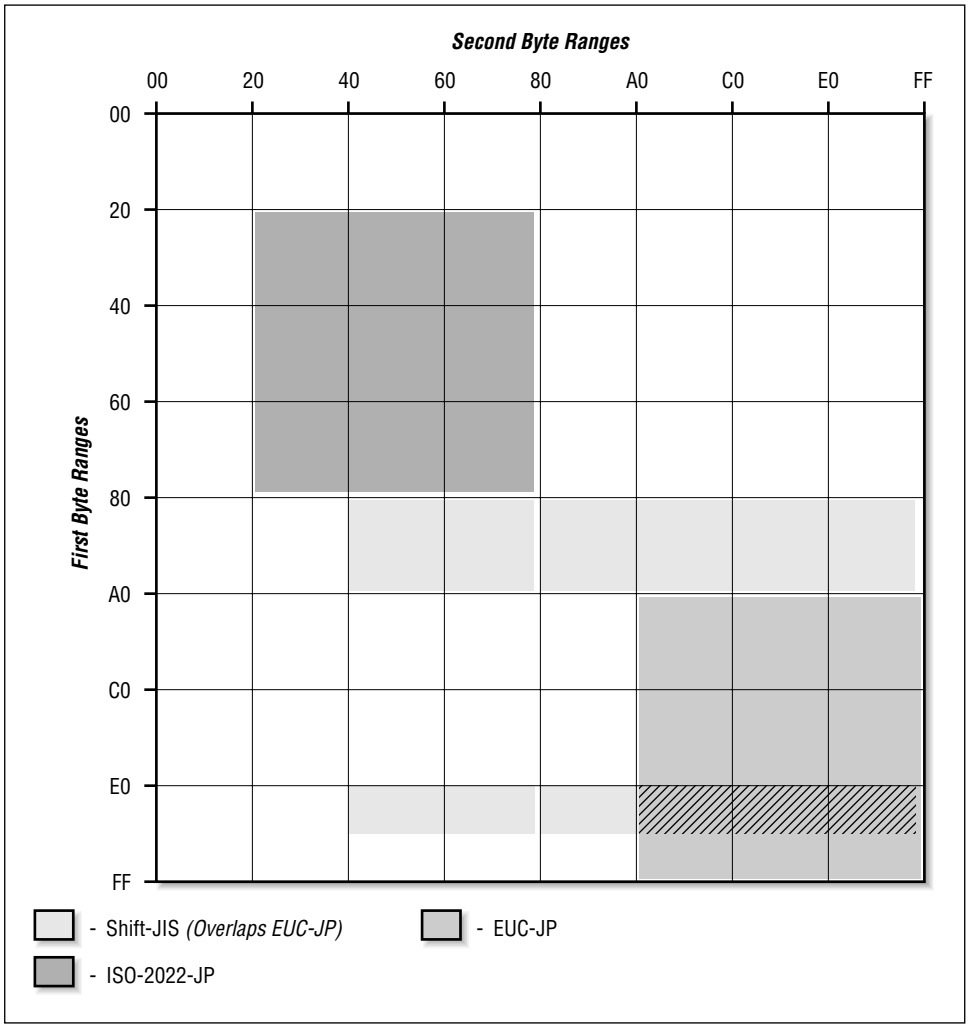
- **GBK (China)**
  - — Expanded EUC-CN encoding.
- **Big Five (Taiwan)**
  - — Disjoint 94×157 matrix (94×63 plus 94×94).
- **Big Five Plus (Taiwan)**
  - — Expanded Big Five encoding.
- **Shift-JIS (Japan)**
  - — Algorithmic derivation of 94×94 into 188-cell rows, skipping 0xA0 through 0xDF (half-width katakana).
- **Johab (Korea)**
  - — Encodes 11,172 hangul as three five-bit units.

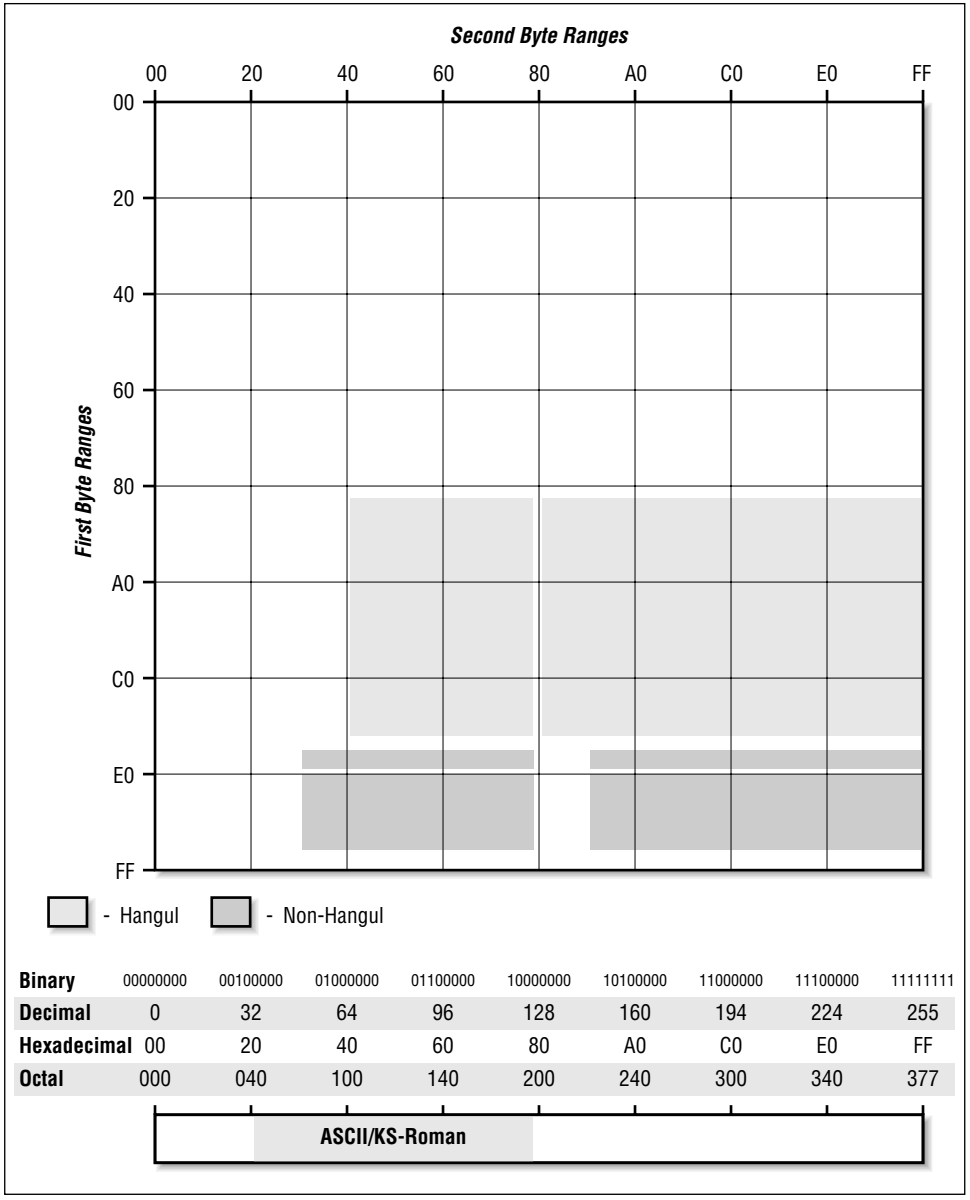

Second Byte Ranges
00
20
40
60
80
A0
C0
E0
FF
First Byte Ranges
00
20
40
60
80
A0
C0
E0
FF
- GBK
- EUC-CN (Overlaps GBK)



Binary	00000000	00100000	01000000	01100000	10000000	10100000	11000000	11100000	11111111
Decimal	0	32	64	96	128	160	194	224	255
Hexadecimal	00	20	40	60	80	A0	C0	E0	FF
Octal	000	040	100	140	200	240	300	340	377


Second Byte Ranges
00
20
40
60
80
A0
C0
E0
FF
First Byte Ranges
00
20
40
60
80
A0
C0
E0
FF
- Big Five Plus
- Big Five (Overlaps Big Five Plus)

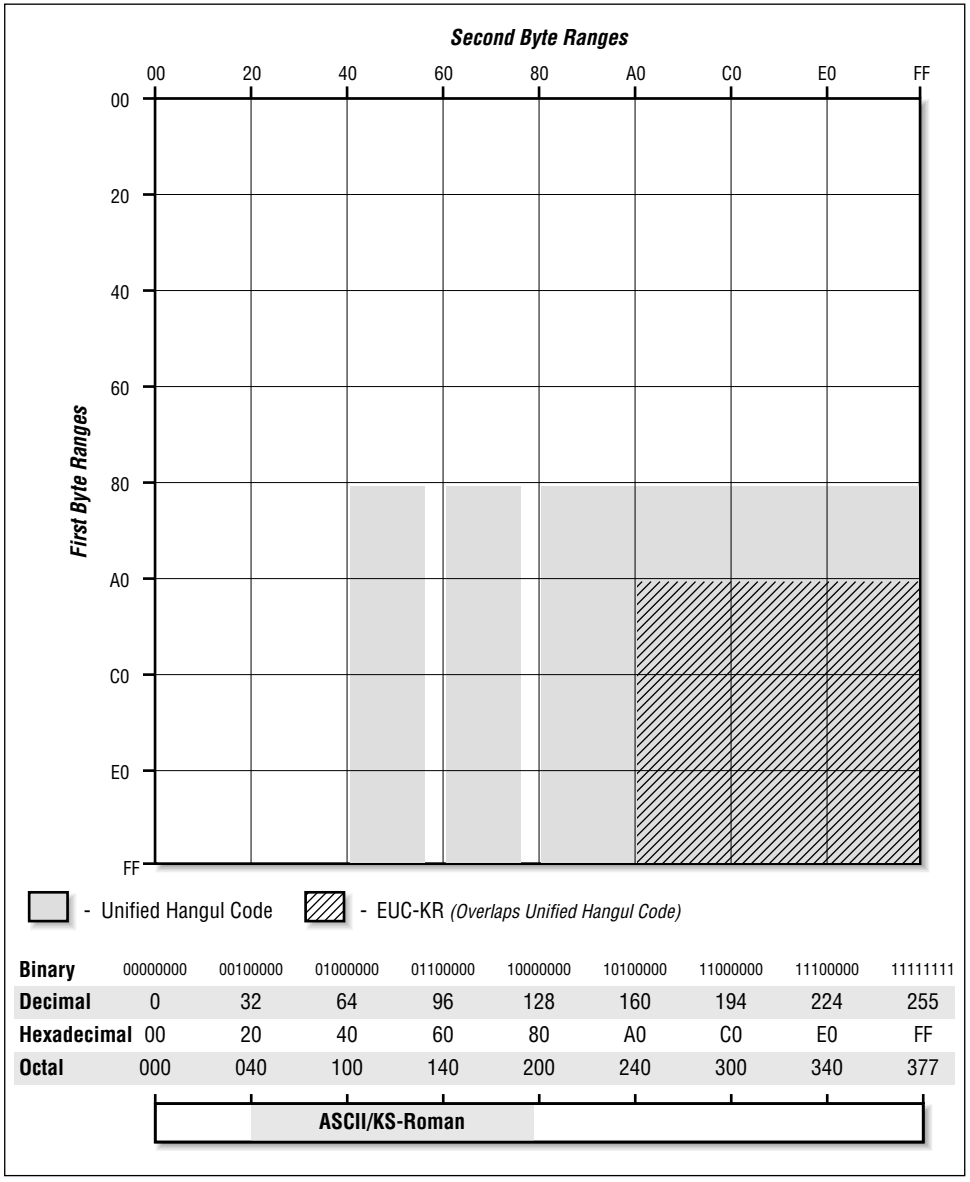

Second Byte Ranges
00
20
40
60
80
A0
C0
E0
FF
First Byte Ranges
00
20
40
60
80
A0
C0
E0
FF
- Shift-JIS (Overlaps EUC-JP)
- EUC-JP
- ISO-2022-JP



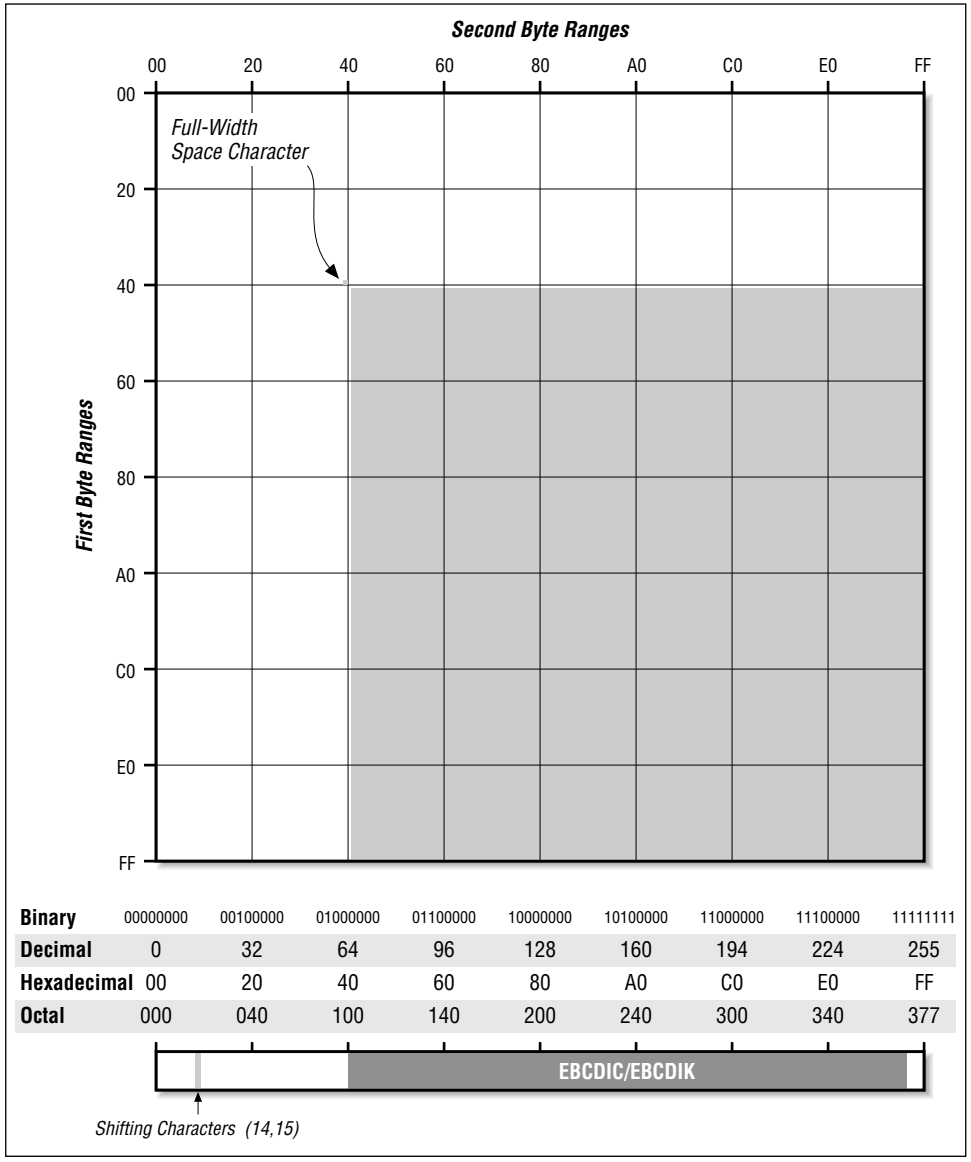
Binary	00000000	00100000	01000000	01100000	10000000	10100000	11000000	11100000	11111111
Decimal	0	32	64	96	128	160	194	224	255
Hexadecimal	00	20	40	60	80	A0	C0	E0	FF
Octal	000	040	100	140	200	240	300	340	377

Vendor-Specific Encodings

- Typically implemented as extensions to standard encodings for MacOS or Windows.
  - — *Example:* MacOS-KH encoding as EUC-KR extension.
  - — *Example:* UHC (Unified Hangul Code) as EUC-KR extension.
- Exception: IBM's encodings.
  - — *Example:* IBM DBCS-Host encoding.
- See Appendixes C and D of *CJKV Information Processing*.



Binary	00000000	00100000	01000000	01100000	10000000	10100000	11000000	11100000	11111111
Decimal	0	32	64	96	128	160	194	224	255
Hexadecimal	00	20	40	60	80	A0	C0	E0	FF
Octal	000	040	100	140	200	240	300	340	377


Second Byte Ranges
00
20
40
60
80
A0
C0
E0
FF
First Byte Ranges
Full-Width
Space Character
Binary 00000000 00100000 01000000 01100000 10000000 10100000 11000000 11100000 11111111
Decimal 0 32 64 96 128 160 194 224 255
Hexadecimal 00 20 40 60 80 A0 C0 E0 FF
Octal 000 040 100 140 200 240 300 340 377
EBCDIC/EBCDIK
Shifting Characters (14,15)

Common Characteristics

- Most "large" character sets combine with a "small" character set—identical or equivalent to ASCII—when encoded according to legacy encodings.
- Most 94×94 character sets have a common structure:
  - Rows 1–15 are reserved for symbols (everything except for Chinese characters and hangul).
  - Rows 16–94 are reserved for Chinese characters or hangul.
  - GB, JIS, KS, and TCVN standards follow this principle.
  - CNS 11643-1992 is an exception, for all planes.
- Most 94×94 character sets can be encoded according to ISO 2022 and EUC.

How Does Unicode Fit In?

- **Supports a large number of legacy character sets.**
  - Full and partial support.
- **Conversion to/from legacy encodings is table-driven.**
  - Not a “bad” thing.
  - Exceptions are ASCII and ISO-8859-1:1998.
- **Available encodings:**
  - UTF-16/UCS-2
  - UTF-8
  - UTF-7
  - Conversion between these encodings is purely algorithmic.
- **Same characters in legacy character sets, but a different encoding.**

Handling Character Variants

- **Many character sets (including vendor-specific) include character variants.**
  - JIS X 0208:1997 includes five variants for 剣 (23-85), specifically 劍 (49-88), 劒 (49-89), 劔 (49-90), 劍 (49-91), and 釼 (78-63).
  - Some planes of CNS 11643-1992 contain character variants.
  - IBM Selected Kanji includes the traditional form of 黒 (25-85), specifically 黑 (0xFC4B)—*available on Windows-J, but not MacOS-J.*

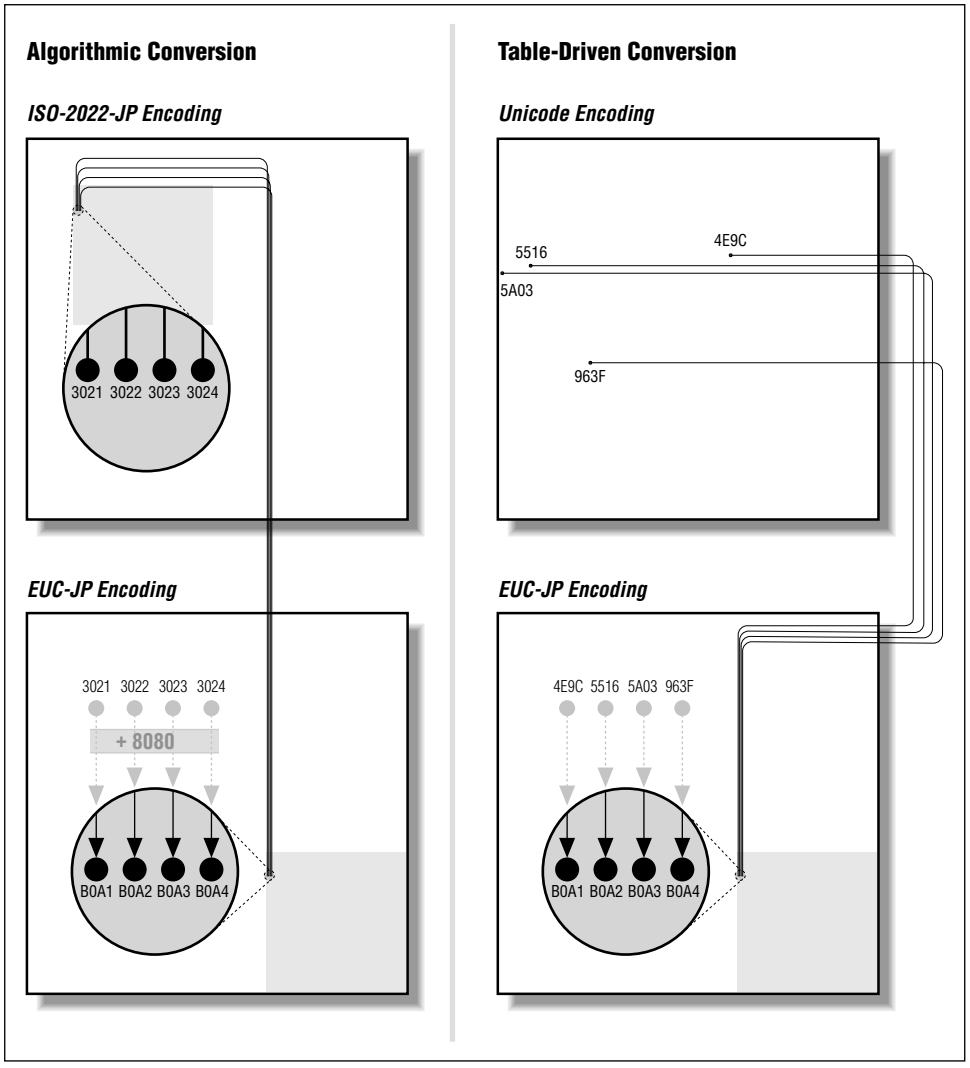
Handling Character Variants (Cont'd)

- **Some character variants are made available in specialized font products—so-called "Gaiji" fonts.**
  - JIS X 0208:1997 includes only two variants for 辺 (42-53), specifically 邊 (78-20) and 邉 (78-21).
  - DTP center Biblos provides the following additional eight variants: 邊 邊 邊 邉 邉 邉 邉 邉.
  - Enfour Media provides the following additional 18 variants: 邊 邊 邊 邊 邊 邊 邊 邊 邊 邉 邉 邉 邉 邉 邉 邉 邉 邉.
- **"Variant tagging" can be a solution.**
  - A proposal has already been submitted.
  - Depends on font availability.

Code Conversion Issues

Adobe

- **Ideographs are not usually a problem.**
  - — JIS C 6226-1978 (JIS78) variants can be problematic due to incomplete coverage in Unicode—some in JIS X 0212-1990.
- **Non-ideographs sometimes cause problems.**
  - — U+2003 or U+3000 for full-width space?
- **Table-driven versus algorithmic code conversion**
  - — Algorithmic code conversion simply applies a set of rules that affect all code points.
  - — Conversion between encodings for a single local is typically algorithmic—ISO-2022-JP, EUC-JP, and Shift-JIS.
  - — Conversion between Unicode and legacy encodings is almost always table-driven.


Algorithmic Conversion
ISO-2022-JP Encoding
3021 3022 3023 3024
EUC-JP Encoding
3021 3022 3023 3024
+ 8080
B0A1 B0A2 B0A3 B0A4
Table-Driven Conversion
Unicode Encoding
5516
4E9C
5A03
963F
EUC-JP Encoding
4E9C 5516 5A03 963F
B0A1 B0A2 B0A3 B0A4

Code Conversion Issues (Cont'd)

- Unicode can act as an effective information interchange code, which has been proven to greatly simplify cross-locale code conversion.
  - Implemented in CJKVConv.pl

    ftp://ftp.oreilly.com/pub/examples/nutshell/cjkv/perl/cjkvconv.pl
  - Implemented in Basis Technology's Uniconv

    http://rosette.basistech.com/demo.html

For More Information

- *The Unicode Standard, Version 2.0* (Addison-Wesley, 1996, ISBN 1-201-48345-9)
  - Version 3.0 due out soon.
- *CJKV Information Processing* (O'Reilly & Associates, 1999, ISBN 1-56592-224-7)
  - Chapters 3 and 4 (character sets and encodings)
  - Appendixes C and D (vendor character sets and encodings)
  - Appendixes E through S (character set and mapping tables)
- National Standards

Q & A

Adobe