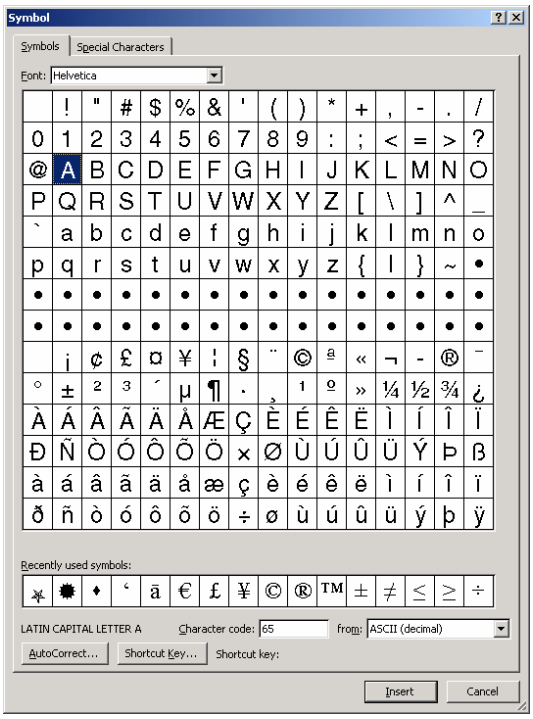
A Quick Explanation of Character Encoding

SimulTrans, L.L.C., +1-650-605-1300, info@simultrans.com, www.simultrans.com

What is character encoding?

Character encoding is the organization of the set of numeric codes that represent all the meaningful characters of a script system in memory. Each character is stored in memory as a number. When a user enters characters, the user's keypresses are converted to character codes; when the characters are displayed onscreen, the character codes are converted to the glyphs of a font. Character encoding is matching the binary representation of a character with the printed character based on a table.



Using the standard "Insert: Symbol" command in Word, we see this table of characters. The group shown here is the "extended ASCII" table of 256 code points.

The character "A" is character 65, out of the 256 code points available in the standard 8-bit ASCII character set.

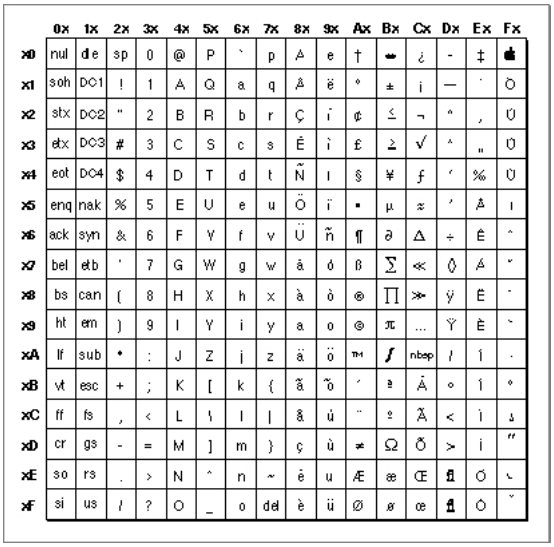
A = $41 hex, ASCII
65 decimal, ASCII
$20 hex, Unicode

7-bit encoding = 2^7 = 128 code points
(32 control characters and 96 printing characters)

A = 1000001

8-bit encoding = 2^8 = 256 code points
(190 printing characters)

The Macintosh uses a slightly different form of extended ASCII, referred to by Apple as "The Standard Roman Character Set." As seen in the table at the right, Apple also usually refers to characters by Hexadecimal placement within the table.



What causes those strange Ó characters in browsers?

> "If the teacher wants to know if a student meets a given standard or not, there should be two levels, ÒyesÓ and ÒnoÓ."

In most character encoding standards, the character set changes to represent the language needed by the user, so the upper-level characters may include symbols, accented Roman letters, Cyrillic, or other characters, depending on the character encoding chosen. For example, the character "Ó" in the Macintosh Standard Roman Character Set is in the same code point as a closing quotation mark in Windows extended ASCII encoding.

References

The Unicode Standard, Version 3.0 by The Unicode Consortium (Editor)
Unicode: A Primer by Tony Graham
CJKV Information Processing by Ken Lunde
Inside Macintosh: Text by Apple Computer (out of print)
Developing International Software by Nadine Kano (out of print)
SimulTrans' website at http://www.simultrans.com

A Few Popular Forms of Character Encoding

SimulTrans, L.L.C., +1-650-605-1300, info@simultrans.com, www.simultrans.com

ISO 646: Based on 7-bit ASCII, with ten character positions for national variants

The International Organization for Standardization first created a standard called **ISO 2022**, which outlines how 7-bit and 8-bit character codes are to be structured and extended.

ISO 8859-1: "Latin 1" encoding, extension of ASCII containing many European characters

ISO 8859 contains encoding standards for various Western and Eastern European character sets:

- Part 1: Latin alphabet No. 1 (Revised 1998)
  Character sets of Western European languages
- Part 2: Latin alphabet No. 2
  Character sets of Eastern European languages (Slavic, Albanian, Hungarian, Romanian)
- Part 3: Latin alphabet No. 3
  Character sets of Southern European languages (Maltese) plus Esperanto
- Part 4: Latin alphabet No. 4 (1998)
  Northern European languages
- Part 5: Latin/Cyrillic alphabet
- Part 6: Latin/Arabic alphabet
- Part 7: Latin/Greek alphabet
- Part 8: Latin/Hebrew alphabet
- Part 9: Latin alphabet No. 5
  Latin character set used for modern Turkish
- Part 10: Latin alphabet No. 6 (1998)
  Icelandic, Nordic, and Baltic character sets
- Part 13 (DIS) Latin alphabet No. 7
- Part 14 (DIS) Latin alphabet No. 8 (Celtic)

The Japanese Industrial Standards Committee has created encoding systems for Japanese text:

JIS X 0201-1976

- numerals (10)
- Latin alphabet (52)
- symbols (32)
- non-printing characters (34)
- katakana (63 half-width characters)

JIS X 0208-1990

- punctuation, symbols (93, 53)
- ISO 646 alphanumerics (10 numerals, 52 characters)
- hiragana (83)
- katakana (86)
- Greek alphabet (48)
- Cyrillic (Russian) alphabet (66)
- line drawing elements (32)
- kanji level 1 (2,965 characters, ordered by Chinese style reading)
- kanji level 2 (3,390 characters, order by Chinese character radical)
- miscellaneous kanji (6 characters)

Microsoft Corporation invented an encoding method for the JIS character set called **Shift-JIS**, which eliminates the escape sequences, and thus the need to switch between character sets.

Simplified Chinese encoding used in Mainland China:

GB 2312-80

- symbols (94)
- numerals (72)
- ISO 646-CN (94 full-width characters)
- hiragana (83)
- katakana (86)
- Greek alphabet (48)
- Cyrillic (Russian) alphabet (66)
- pinyin and bopomofo characters (26, 37)
- line-drawing elements (76)
- hanzi level 1 (3,755, ordered by pinyin reading)
- hanzi level 2 (3,008, ordered by Chinese character radical, then stroke)

Simplified Chinese uses the following encoding methods: 7-bit ISO 2022, ISO-2022-CN (e-mail message encoding), EUC-CN, and HZ (HZ-GB-2312).

Traditional Chinese encoding used in Taiwan:

Big-5

- symbols (157)
- symbols (157)
- symbols (94)
- hanzi level 1 (5,401 Chinese characters)
- hanzi level 2 (7,652 Chinese characters)
  (Characters are ordered by number of strokes, then radical.)

CNS 11643-1992

- symbols (438)
- classical radicals (213)
- graphic representations of control characters (33)
- hanzi 1 (5,401 Chinese characters)
- hanzi 2 (7,650 Chinese characters)
- hanzi 3 (6,148 Chinese characters)
- hanzi 4 (7,298 Chinese characters)
- hanzi 5 (8,603 Chinese characters)
- hanzi 6 (6,388 Chinese characters)
- hanzi 7 (6,539 Chinese characters)

Traditional Chinese is encoded with the following methods: 7-bit ISO 2022, ISO-2022-CN (e-mail message encoding), EUC-TW, and Big-5. Big-5 is the encoding system traditionally used on both Windows and Macintosh operating systems. EUC is primarily used by UNIX.

The all-encompassing Unicode:

Unicode (ISO 10646-1: 1993)

- ISO 646
- ISO 8859-1
  Eastern European accented characters
- International Phonetic Alphabet (IPA)
- Greek (including accented characters)
- Cyrillic, Georgian and Armenian
- Hebrew
- Arabic characters (all four forms)
- Indian subcontinent character sets
- Thai and Lao
- Chinese/Japanese/Korean (CJK) ideographic characters
- Mathematical operators and special character forms
- Box and line drawing characters
- Geometric shapes and Dingbats
- Special OCR characters used on checks
- Encircled characters and numbers