

# Computer Science Basics

## Encoding

Fall Term 2025/2026

E. Benoist with H. Wenger & C. Grothoff | BFH-TI

# Encoding

- ▶ Last Week
- ▶ Background
- ▶ Typography
- ▶ History
- ▶ Unicode
- ▶ UTF-8
- ▶ Conclusion

# Last Week

# Linux and file types

- **Unix is representing 95% of all servers**
  - GNU/Linux, Solaris, FreeBSD, OpenBSD, MacOS-X, ...
- **But also 100% of all Smartphones**
  - Android and iOS are both Unix systems.
- **Different types of files : Text files**
  - stored in ASCII or other encodings (topic of next week)
  - Source of programs, texts, XML, HTML, CSS,
- **Different types of files : Binary files**
  - Write data with a specific encoding
  - Programs (runnable), Databases, ...
  - Some standards are proprietary some are open.
  - We will see standards for images.

# Background

## How do computers represent characters?

- Glyphs, ligatures and fonts
- Historical standards
- Unicode 5 layer architecture:
  1. Abstract Character Repertoire
  2. Coded Character Set
  3. Character Encoding Form
  4. Character Encoding Scheme
  5. Transfer Encoding Syntax

# Typography

# Glyph



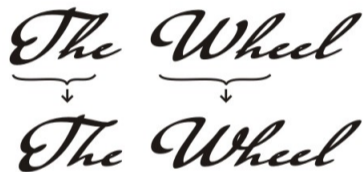
“A grapheme (computing: character) is the smallest unit of a writing system of any given language.”

–[https://en.](https://en.wikipedia.org/wiki/Grapheme)

[wikipedia.org/wiki/Grapheme](https://en.wikipedia.org/wiki/Grapheme)

“A glyph (computing: shape) is an elemental symbol within an agreed set of symbols, intended to represent a readable character for the purposes

# Ligature



A *ligature* is the combination of two or more graphemes (or letters) into a single glyph.

# Font

*A font defines size, weight and style of a **typeface**.*

*A typeface is a set of glyphs that share common design features.*

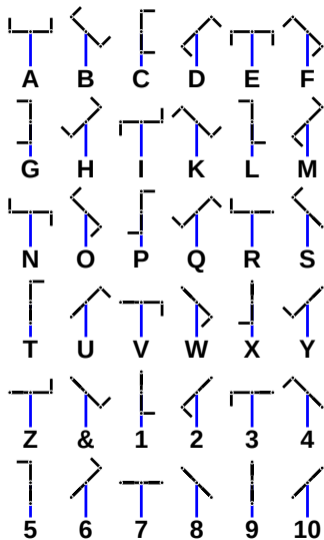
A font may not define a unique glyph for every character. For example, the Latin and Greek 'A' may be different characters sharing the same glyph.

# Character

- Unit of information
- Used for organization, control or representation of textual data

# History

# Telegraph (Chappe, 1792)



# Extended Binary Coded Decimal Interchange Code (1963)

**EBCDIC character codes**

1st hex digit

2nd hex digit

|   | 0   | 1   | 2   | 3   | 4         | 5  | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|---|-----|-----|-----|-----|-----------|----|---|---|---|---|---|---|---|---|---|---|
| 0 | NUL | DLE | DS  |     | SP        | &  | - |   |   |   |   |   |   |   |   | 0 |
| 1 | SOH | DC1 | SOS |     |           |    | / |   | a | j |   |   | A | J |   | 1 |
| 2 | STX | DC2 | FS  | SYN |           |    |   |   | b | k | s |   | B | K | S | 2 |
| 3 | ETX | TM  |     |     |           |    |   |   | c | l | t |   | C | L | T | 3 |
| 4 | PF  | RES | BYP | PN  |           |    |   |   | d | m | u |   | D | M | U | 4 |
| 5 | HT  | NL  | LF  | RS  |           |    |   |   | e | n | v |   | E | N | V | 5 |
| 6 | LC  | BS  | ETB | UC  |           |    |   |   | f | o | w |   | F | O | W | 6 |
| 7 | DEL | IL  | ESC | EOT |           |    |   |   | g | p | x |   | G | P | X | 7 |
| 8 |     | CAN |     |     |           |    |   |   | h | q | y |   | H | Q | Y | 8 |
| 9 |     | EM  |     |     |           |    |   |   | i | r | z |   | I | R | Z | 9 |
| A | SMM | CC  | SM  |     | C<br>CENT | !  | : |   |   |   |   |   |   |   |   |   |
| B | VT  | UB  | CB  |     | US        | \$ | # |   |   |   |   |   |   |   |   |   |
| C | FF  | WS  | DC4 |     | <         | *  | % | @ |   |   |   |   |   |   |   |   |

# ASCII TABLE

| Decimal | Hexadecimal | Binary | Octal | Char                   | Decimal | Hexadecimal | Binary  | Octal | Char | Decimal | Hexadecimal | Binary  | Octal | Char  |
|---------|-------------|--------|-------|------------------------|---------|-------------|---------|-------|------|---------|-------------|---------|-------|-------|
| 0       | 0           | 0      | 0     | [NULL]                 | 48      | 30          | 110000  | 60    | 0    | 96      | 60          | 1100000 | 140   | `     |
| 1       | 1           | 1      | 1     | [START OF HEADING]     | 49      | 31          | 110001  | 61    | 1    | 97      | 61          | 1100001 | 141   | a     |
| 2       | 2           | 10     | 2     | [START OF TEXT]        | 50      | 32          | 110010  | 62    | 2    | 98      | 62          | 1100010 | 142   | b     |
| 3       | 3           | 11     | 3     | [END OF TEXT]          | 51      | 33          | 110011  | 63    | 3    | 99      | 63          | 1100011 | 143   | c     |
| 4       | 4           | 100    | 4     | [END OF TRANSMISSION]  | 52      | 34          | 110100  | 64    | 4    | 100     | 64          | 1100100 | 144   | d     |
| 5       | 5           | 101    | 5     | [ENQUIRY]              | 53      | 35          | 110101  | 65    | 5    | 101     | 65          | 1100101 | 145   | e     |
| 6       | 6           | 110    | 6     | [ACKNOWLEDGE]          | 54      | 36          | 110110  | 66    | 6    | 102     | 66          | 1100110 | 146   | f     |
| 7       | 7           | 111    | 7     | [BELL]                 | 55      | 37          | 110111  | 67    | 7    | 103     | 67          | 1100111 | 147   | g     |
| 8       | 8           | 1000   | 10    | [BACKSPACE]            | 56      | 38          | 111000  | 70    | 8    | 104     | 68          | 1101000 | 150   | h     |
| 9       | 9           | 1001   | 11    | [HORIZONTAL TAB]       | 57      | 39          | 111001  | 71    | 9    | 105     | 69          | 1101001 | 151   | i     |
| 10      | A           | 1010   | 12    | [LINE FEED]            | 58      | 3A          | 111010  | 72    | :    | 106     | 6A          | 1101010 | 152   | j     |
| 11      | B           | 1011   | 13    | [VERTICAL TAB]         | 59      | 3B          | 111011  | 73    | ;    | 107     | 6B          | 1101011 | 153   | k     |
| 12      | C           | 1100   | 14    | [FORM FEED]            | 60      | 3C          | 111100  | 74    | <    | 108     | 6C          | 1101100 | 154   | l     |
| 13      | D           | 1101   | 15    | [CARRIAGE RETURN]      | 61      | 3D          | 111101  | 75    | =    | 109     | 6D          | 1101101 | 155   | m     |
| 14      | E           | 1110   | 16    | [SHIFT OUT]            | 62      | 3E          | 111110  | 76    | >    | 110     | 6E          | 1101110 | 156   | n     |
| 15      | F           | 1111   | 17    | [SHIFT IN]             | 63      | 3F          | 111111  | 77    | ?    | 111     | 6F          | 1101111 | 157   | o     |
| 16      | 10          | 10000  | 20    | [DATA LINK ESCAPE]     | 64      | 40          | 1000000 | 100   | @    | 112     | 70          | 1110000 | 160   | p     |
| 17      | 11          | 10001  | 21    | [DEVICE CONTROL 1]     | 65      | 41          | 1000001 | 101   | A    | 113     | 71          | 1110001 | 161   | q     |
| 18      | 12          | 10010  | 22    | [DEVICE CONTROL 2]     | 66      | 42          | 1000010 | 102   | B    | 114     | 72          | 1110010 | 162   | r     |
| 19      | 13          | 10011  | 23    | [DEVICE CONTROL 3]     | 67      | 43          | 1000011 | 103   | C    | 115     | 73          | 1110011 | 163   | s     |
| 20      | 14          | 10100  | 24    | [DEVICE CONTROL 4]     | 68      | 44          | 1000100 | 104   | D    | 116     | 74          | 1110100 | 164   | t     |
| 21      | 15          | 10101  | 25    | [NEGATIVE ACKNOWLEDGE] | 69      | 45          | 1000101 | 105   | E    | 117     | 75          | 1110101 | 165   | u     |
| 22      | 16          | 10110  | 26    | [SYNCHRONOUS IDLE]     | 70      | 46          | 1000110 | 106   | F    | 118     | 76          | 1110110 | 166   | v     |
| 23      | 17          | 10111  | 27    | [ENG OF TRANS. BLOCK]  | 71      | 47          | 1000111 | 107   | G    | 119     | 77          | 1110111 | 167   | w     |
| 24      | 18          | 11000  | 30    | [CANCEL]               | 72      | 48          | 1001000 | 110   | H    | 120     | 78          | 1111000 | 170   | x     |
| 25      | 19          | 11001  | 31    | [END OF MEDIUM]        | 73      | 49          | 1001001 | 111   | I    | 121     | 79          | 1111001 | 171   | y     |
| 26      | 1A          | 11010  | 32    | [SUBSTITUTE]           | 74      | 4A          | 1001010 | 112   | J    | 122     | 7A          | 1111010 | 172   | z     |
| 27      | 1B          | 11011  | 33    | [ESCAPE]               | 75      | 4B          | 1001011 | 113   | K    | 123     | 7B          | 1111011 | 173   | {     |
| 28      | 1C          | 11100  | 34    | [FILE SEPARATOR]       | 76      | 4C          | 1001100 | 114   | L    | 124     | 7C          | 1111100 | 174   |       |
| 29      | 1D          | 11101  | 35    | [GROUP SEPARATOR]      | 77      | 4D          | 1001101 | 115   | M    | 125     | 7D          | 1111101 | 175   | }     |
| 30      | 1E          | 11110  | 36    | [RECORD SEPARATOR]     | 78      | 4E          | 1001110 | 116   | N    | 126     | 7E          | 1111110 | 176   | ~     |
| 31      | 1F          | 11111  | 37    | [UNIT SEPARATOR]       | 79      | 4F          | 1001111 | 117   | O    | 127     | 7F          | 1111111 | 177   | [DEL] |
| 32      | 20          | 100000 | 40    | [SPACE]                | 80      | 50          | 1010000 | 120   | P    |         |             |         |       |       |
| 33      | 21          | 100001 | 41    | !                      | 81      | 51          | 1010001 | 121   | Q    |         |             |         |       |       |
| 34      | 22          | 100010 | 42    | "                      | 82      | 52          | 1010010 | 122   | R    |         |             |         |       |       |
| 35      | 23          | 100011 | 43    | #                      | 83      | 53          | 1010011 | 123   | S    |         |             |         |       |       |
| 36      | 24          | 100100 | 44    | \$                     | 84      | 54          | 1010100 | 124   | T    |         |             |         |       |       |
| 37      | 25          | 100101 | 45    | %                      | 85      | 55          | 1010101 | 125   | U    |         |             |         |       |       |
| 38      | 26          | 100110 | 46    | &                      | 86      | 56          | 1010110 | 126   | V    |         |             |         |       |       |

# Control codes

Character sets often contain control codes, code positions that are not mapped to a visible character but for control. Examples:

**CR** Carriage return

**LF** Line feed

**HT** Horizontal tab

**CTRL-C** Interrupt

**CTRL-D** End of input

End of line on Mac is CR, on Unix LF and on Windows CR LF.

# Control



# ISO 8859-1: Extended ASCII (latin1)

|    |    |    |    |   |    |   |    |    |    |    |    |   |   |   |   |
|----|----|----|----|---|----|---|----|----|----|----|----|---|---|---|---|
|    | A1 | A2 | A3 |   | A5 |   | A7 | A8 | A9 | AA | AB |   |   |   |   |
|    | ı  | ϕ  | £  |   | ¥  |   | §  | ℥  | ©  | ™  | «  |   |   |   |   |
| B0 | ±  | ²  | ³  |   | μ  | ¶ | •  |    | ¹  | º  | »  | ¼ | ½ |   | ¿ |
| C0 | À  | Á  | Â  | Ã | Ä  | Å | Ç  | È  | É  | Ê  | Ë  | Ì | Í | Î | Ï |
|    | Ñ  | Ò  | Ó  | Ô | Õ  | Ö | Ø  | Ù  | Ú  | Û  | Ü  | Ý |   |   | ß |
| E0 | à  | á  | â  | ã | ä  | å | ç  | è  | é  | ê  | ë  | ì | í | î | ï |
|    | ñ  | ò  | ó  | ô | õ  | ö | ø  | ù  | ú  | û  | ü  | ý |   |   |   |

# CP1252: Extended ASCII (winlatin1)

|    |    |    |    |    |    |    |    |    |     |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |   |
|----|----|----|----|----|----|----|----|----|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|---|
| 80 | €  | 82 | ,  | 83 | f  | 84 | ,, | 85 | ... | 86 | †  | ‡  | 88 | ˆ  | 89 | %  | 9A | Š  | 9B | <  | 9C | Œ  | 9E | Ž  |    |    |    |    |    |    |   |
|    | 91 | ç  | 92 | ,  | 93 | “  | 94 | ”  | 95  | •  | 96 | -  | 97 | -  | 98 | ˜  | 99 | ™  | 9A | Š  | 9B | >  | 9C | œ  | 9E | ž  | 9F | ÿ  |    |    |   |
| A0 | A1 | ı  | A2 | ϕ  | A3 | £  | A4 | ℥  | A5  | ¥  | A6 | ı  | A7 | §  | A8 | .. | A9 | ©  | AA | ı  | AB | «  | AC | ¬  | AD | -  | AE | ®  | AF | -  |   |
| B0 | °  | B1 | ±  | B2 | ²  | B3 | ³  | B4 | ˘   | B5 | μ  | B6 | ¶  | B7 | ·  | B8 | ˙  | B9 | ¹  | BA | º  | BB | »  | BC | ¼  | BD | ½  | BE | ¾  | BF | ¿ |
| C0 | À  | C1 | Á  | C2 | Â  | C3 | Ã  | C4 | Ä   | C5 | Å  | C6 | Æ  | C7 | Ç  | C8 | È  | C9 | É  | CA | Ê  | CB | Ë  | CC | Ì  | CD | Í  | CE | Î  | CF | Ï |
| D0 | Ð  | D1 | Ñ  | D2 | Ò  | D3 | Ó  | D4 | Ô   | D5 | Õ  | D6 | Ö  | D7 | ×  | D8 | Ø  | D9 | Ù  | DA | Ú  | DB | Û  | DC | Ü  | DD | Ý  | DE | Þ  | DF | ß |
| E0 | à  | E1 | á  | E2 | â  | E3 | ã  | E4 | ä   | E5 | å  | E6 | æ  | E7 | ç  | E8 | è  | E9 | é  | EA | ê  | EB | ë  | EC | ì  | ED | í  | EE | î  | EF | ï |
| F0 | ä  | F1 | ñ  | F2 | ò  | F3 | ó  | F4 | ô   | F5 | õ  | F6 | ö  | F7 | ÷  | F8 | ø  | F9 | ù  | FA | ú  | FB | û  | FC | ü  | FD | ý  | FE | þ  | FF | ÿ |

# ISO 8859-2: Extended ASCII (latin2)

|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A0 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | AA | AB | AC | AD | AE | AF |
|    | Ā  | Ĳ  | Ł  | Ĥ  | Ĺ  | Š  | Ś  | ŀ  | Š  | Ş  | Ť  | Ž  | -  | Ž  | Ž  |
| B0 | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | BA | BB | BC | BD | BE | BF |
| ◊  | ā  | ĳ  | ł  | ˆ  | ĺ  | š  | ś  | ˚  | š  | ş  | ť  | ž  | ˘  | ž  | ž  |
| C0 | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | CA | CB | CC | CD | CE | CF |
| Ř  | Ā  | Ā  | Ā  | Ā  | Ĺ  | Č  | Ç  | Č  | Ě  | Ě  | Ě  | Ě  | Ī  | Ī  | Ď  |
| D0 | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | DA | DB | DC | DD | DE | DF |
| Đ  | Ñ  | Ñ  | Ō  | Ô  | Ŏ  | Ö  | ×  | Ř  | Ů  | Ú  | Ů  | Ü  | Ÿ  | Ť  | ß  |
| E0 | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | EA | EB | EC | ED | EE | EF |
| ř  | ā  | ā  | ā  | ā  | ĺ  | č  | ç  | č  | ě  | ě  | ě  | ě  | ī  | ī  | ď  |
| F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | FA | FB | FC | FD | FE | FF |
| đ  | ñ  | ñ  | ō  | ô  | ǎ  | ö  | ÷  | ř  | ů  | ú  | ů  | ü  | ý  | ţ  | .  |

# ISO 8859-3: Extended ASCII (latin3)

|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A0 | A1 | A2 | A3 | A4 |    | A6 | A7 | A8 | A9 | AA | AB | AC | AD |    | AF |
|    | Ħ  | ı  | £  | ℥  |    | Ĥ  | §  | ¨  | İ  | Ş  | Ğ  | Ĵ  | -  |    | Ž  |
| B0 | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | BA | BB | BC | BD |    | BF |
| °  | ħ  | 2  | 3  | ˘  | ı  | ĥ  | •  | ,  | ı  | Ş  | ğ  | ĵ  | ¼  |    | ž  |
| C0 | C1 | C2 |    | C4 | C5 | C6 | C7 | C8 | C9 | CA | CB | CC | CD | CE | CF |
| À  | Á  | Â  |    | Ä  | Å  | Ç  | Ç  | È  | É  | Ê  | Ë  | Ï  | Ï  | Î  | Ï  |
|    | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | DA | DB | DC | DD | DE | DF |
|    | Ñ  | Ò  | Ó  | Ô  | Õ  | Ö  | ×  | Ğ  | Û  | Ü  | Û  | Ü  | Ü  | Ŝ  | ß  |
| E0 | E1 | E2 |    | E4 | E5 | E6 | E7 | E8 | E9 | EA | EB | EC | ED | EE | EF |
| à  | á  | â  |    | ä  | å  | ç  | ç  | è  | é  | ê  | ë  | ï  | ï  | î  | ï  |
|    | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | FA | FB | FC | FD | FE | FF |
|    | ñ  | ò  | ó  | ô  | õ  | ö  | ÷  | ğ  | û  | ü  | ü  | ü  | ü  | ŝ  | •  |

# ISO 8859-4: Extended ASCII (latin4)

|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A0 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | AA | AB | AC | AD | AE | AF |
|    | À  | Ā  | Ŕ  | Ɑ  | İ  | Ľ  | Š  | .. | Š  | Ě  | Ǫ  | Ʀ  | -  | Ž  | -  |
| B0 | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | BA | BB | BC | BD | BE | BF |
| ◊  | ą  | ˆ  | ŕ  | ˘  | ĩ  | ł  | ˘  | ˘  | š  | ě  | ǫ  | ř  | ŋ  | ž  | ŋ  |
| C0 | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | CA | CB | CC | CD | CE | CF |
| Ā  | Ă  | Â  | Ã  | Ä  | Å  | Æ  | Į  | Č  | Ě  | Ɛ  | Ë  | È  | Ì  | Î  | Ī  |
| D0 | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | DA | DB | DC | DD | DE | DF |
| Đ  | Ń  | Ō  | Ɔ  | Ô  | Õ  | Ö  | ×  | Ø  | Ų  | Ú  | Û  | Ü  | Û  | Ū  | Ɔ  |
| E0 | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | EA | EB | EC | ED | EE | EF |
| ā  | ă  | â  | ã  | ä  | å  | æ  | į  | č  | ě  | ɛ  | ë  | è  | ì  | î  | ī  |
| F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | FA | FB | FC | FD | FE | FF |
| đ  | ń  | ō  | ƙ  | ô  | õ  | ö  | ÷  | ø  | ų  | ú  | û  | ü  | Û  | Ū  | ˙  |

## ISO 8859-5: Extended ASCII (Cyrillic)

|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A0 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | AA | AB | AC | AD | AE | AF |
|    | Ё  | Ђ  | Ѓ  | Є  | Ѕ  | І  | Ї  | Ј  | Љ  | Њ  | Ћ  | Ќ  | -  | Ў  | Ў  |
| B0 | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | BA | BB | BC | BD | BE | BF |
| А  | Б  | В  | Г  | Д  | Е  | Ж  | З  | И  | Й  | К  | Л  | М  | Н  | О  | П  |
| C0 | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | CA | CB | CC | CD | CE | CF |
| Р  | С  | Т  | У  | Ф  | Х  | Ц  | Ч  | Ш  | Щ  | Ъ  | Ы  | Ь  | Э  | Ю  | Я  |
| D0 | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | DA | DB | DC | DD | DE | DF |
| а  | б  | в  | г  | д  | е  | ж  | з  | и  | й  | к  | л  | м  | н  | о  | п  |
| E0 | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | EA | EB | EC | ED | EE | EF |
| р  | с  | т  | у  | ф  | х  | ц  | ч  | ш  | щ  | ъ  | ы  | ь  | э  | ю  | я  |
| F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | FA | FB | FC | FD | FE | FF |
| №  | ё  | ђ  | ѓ  | є  | ѕ  | і  | ї  | ј  | љ  | њ  | ќ  | ќ  | ѕ  | ў  | џ  |

# ISO 8859-6: Extended ASCII (Arabic)

|    |    |    |    |     |    |    |    |    |    |    |    |    |    |    |    |    |    |   |
|----|----|----|----|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|---|
| A0 |    |    |    | A4  | ٨  |    |    |    |    |    |    | AC | ء  | AD | ـ  |    |    |   |
|    |    |    |    |     |    |    |    |    |    |    |    | BB | ة  |    |    |    | BF | ؟ |
|    | C1 | C2 | C3 | C4  | C5 | C6 | C7 | C8 | C9 | CA | CA | CB | CC | CD | CE | CF |    |   |
|    | ء  | آ  | أ  | ؤ   | اء | سي | ا  | ب  | ة  | ت  | ث  | ج  | ح  | خ  | د  |    |    |   |
| D0 | D1 | D2 | D3 | D4  | D5 | D6 | D7 | D8 | D9 | DA |    |    |    |    |    |    |    |   |
| ذ  | ر  | ز  | سر | نذر | صر | ظ  | ط  | ظ  | ع  | غ  |    |    |    |    |    |    |    |   |
| E0 | E1 | E2 | E3 | E4  | E5 | E6 | E7 | E8 | E9 | EA | EB | EC | ED | EE | EF |    |    |   |
| ـ  | ف  | ق  | ك  | ل   | م  | ن  | ه  | و  | ي  | ي  | ٴ  | ٴ  | ٴ  | ٴ  | ٴ  |    |    |   |
| F0 | F1 | F2 |    |     |    |    |    |    |    |    |    |    |    |    |    |    |    |   |
| ،  | ء  | ه  |    |     |    |    |    |    |    |    |    |    |    |    |    |    |    |   |

# ISO 8859-7: Extended ASCII (Greek)

|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| A0   | A1 ¸ | A2 ‚ | A3 £ |      |      | A6 ¡ | A7 § | A8 ¨ | A9 © |      | AB « | AC ¬ | AD - | AF - |      |
| B0 ° | B1 ± | B2 ² | B3 ³ | B4 ´ | B5 ˆ | B6 ˆ | B7 · | B8 È | B9 Ì | BA Ì | BB » | BC Ò | BD ¼ | BE ÿ | BF ð |
| C0 ì | C1 Å | C2 Β | C3 Γ | C4 Δ | C5 Ε | C6 Ζ | C7 Η | C8 Θ | C9 Ι | CA Κ | CB Λ | CC Μ | CD Ν | CE Ξ | CF Ο |
| D0 Π | D1 Ρ |      | D3 Σ | D4 Τ | D5 Υ | D6 Φ | D7 Χ | D8 Ψ | D9 Ω | DA Ì | DB ÿ | DC Ò | DD È | DE Ñ | DF ì |
| E0 Ù | E1 α | E2 β | E3 γ | E4 δ | E5 ε | E6 ζ | E7 η | E8 θ | E9 ι | EA κ | EB λ | EC μ | ED ν | EE ξ | EF ο |
| F0 π | F1 ρ | F2 ς | F3 σ | F4 τ | F5 υ | F6 φ | F7 χ | F8 ψ | F9 ω | FA ì | FB ÿ | FC Ò | FD Ù | FE ð |      |



# ISO 8859-9: Extended ASCII (latin5)

|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A0 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | AA | AB | AC | AD | AE | AF |
|    | ı  | ϕ  | £  | ℥  | ¥  | ı  | §  | .. | ©  | ı  | «  | ¬  | -  | ®  | -  |
| B0 | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | BA | BB | BC | BD | BE | BF |
| °  | ±  | ²  | ³  | ˘  | ℥  | ¶  | •  | ,  | ¹  | ı  | »  | ¼  | ½  | ¾  | ¿  |
| C0 | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | CA | CB | CC | CD | CE | CF |
| À  | Á  | Â  | Ã  | Ä  | Å  | Æ  | Ç  | È  | É  | Ê  | Ë  | Ì  | Í  | Î  | Ï  |
| D0 | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | DA | DB | DC | DD | DE | DF |
| Ġ  | Ñ  | Ō  | Ȯ  | Ȫ  | Ȭ  | Ö  | ×  | Ø  | Ù  | Ú  | Û  | Ü  | İ  | Ş  | ß  |
| E0 | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | EA | EB | EC | ED | EE | EF |
| Ǻ  | ǻ  | Ǽ  | Ǿ  | ǿ  | Ǻ  | æ  | ç  | è  | é  | ê  | ë  | ì  | í  | î  | ï  |
| F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | FA | FB | FC | FD | FE | FF |
| ǻ  | ñ  | Ȯ  | Ȯ  | Ȫ  | Ȭ  | ö  | ÷  | ø  | ù  | ú  | û  | ü  | ı  | ş  | ÿ  |

# ISO 8859-10: Extended ASCII (latin6)

|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A0 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | AA | AB | AC | AD | AE | AF |
|    | À  | Ā  | Ġ  | Ī  | Ĩ  | Ɔ  | Š  | Ł  | Đ  | Ṧ | Ʀ  | Ž  | -  | Ū  | Ŋ  |
| B0 | á  | ē  | ġ  | ī  | ĩ  | ƀ  | ·  | ł  | đ  | ṧ | Ʀ  | ž  | -  | ū  | ŋ  |
| C0 | Ā  | Ā  | Ā  | Ā  | Ā  | Æ  | ı  | č  | ě  | ĕ  | ë  | é  | í  | î  | ï  |
| D0 | Đ  | Ń  | Ō  | Ȯ  | Ȱ  | Ö  | Ū  | Ø  | Ų  | Ŭ  | Ũ  | Ü  | Ý  | Ɔ  | Ɔ  |
| E0 | ā  | ā  | ā  | ā  | ā  | æ  | ı  | č  | ě  | ĕ  | ë  | é  | í  | î  | ï  |
| F0 | ǎ  | ŋ  | ō  | ȯ  | ȱ  | ö  | ū  | ø  | ų  | ŭ  | ũ  | ü  | ý  | Ɔ  | ƀ  |

# ISO 8859-11: Extended ASCII (Thai)

|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
|    | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | AA | AB | AC | AD | AE | AF |
|    | ก  | ข  | ฃ  | ค  | ฅ  | ฉ  | ช  | ฌ  | ฉ  | ซ  | ฅ  | ฌ  | ญ  | ฉ  | ฉ  |
| B0 | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | BA | BB | BC | BD | BE | BF |
| ๓  | ก  | ข  | ฃ  | ค  | ฅ  | ฉ  | ช  | ฌ  | ฉ  | ซ  | ฅ  | ฌ  | ญ  | ฉ  | ฉ  |
| C0 | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | CA | CB | CC | CD | CE | CF |
| ก  | ข  | ฃ  | ค  | ฅ  | ฉ  | ช  | ฌ  | ฌ  | ฌ  | ฌ  | ฌ  | ฌ  | ฌ  | ฌ  | ก  |
| D0 | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | DA |    |    |    |    | DF |
| ๔  | ๕  | ๖  | ๗  | ๘  | ๙  | ๑  | ๒  | ๓  | ๔  | ๕  |    |    |    |    | ฿  |
| E0 | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | EA | EB | EC | ED | EE | EF |
| ๕  | ๖  | ๗  | ๘  | ๙  | ๑  | ๒  | ๓  | ๔  | ๕  | ๖  | ๗  | ๘  | ๙  | ๑  | ๒  |
| F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | FA | FB |    |    |    |    |
| ๑  | ๒  | ๓  | ๔  | ๕  | ๖  | ๗  | ๘  | ๙  | ๑  | ๒  | ๓  |    |    |    |    |

# ISO 8859-13: Extended ASCII (latin7)

|      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| A0   | A1 „ | A2 Œ | A3 £ | A4 Ɔ | A5 „ | A6 ¡ | A7 § | A8 Ø | A9 © | AA Œ | AB « | AC ¬ | AD - | AE ® | AF Æ |
| B0 ° | B1 ± | B2 ² | B3 ³ | B4 ¨ | B5 µ | B6 ¶ | B7 · | B8 ø | B9 ¹ | BA Œ | BB » | BC ¼ | BD ½ | BE ¾ | BF æ |
| C0 Ā | C1 Ī | C2 Ā | C3 Ć | C4 Ä | C5 Å | C6 Ę | C7 Ē | C8 Ć | C9 Ě | CA Ź | CB Ę | CC Ğ | CD Ķ | CE Ī | CF Ĺ |
| D0 Š | D1 Ń | D2 Ņ | D3 Ő | D4 Ō | D5 Ū | D6 Ö | D7 × | D8 Ų | D9 Ł | DA Ś | DB Ū | DC Ü | DD Ż | DE Ž | DF Ɔ |
| E0 ą | E1 ı | E2 ā | E3 Ć | E4 ä | E5 å | E6 ę | E7 ē | E8 Ć | E9 ě | EA Ź | EB Ę | EC ğ | ED Ķ | EE Ī | EF Ĺ |
| F0 š | F1 ń | F2 ņ | F3 ő | F4 ō | F5 ū | F6 ö | F7 ÷ | F8 ų | F9 ł | FA ś | FB ū | FC ü | FD ż | FE ž | FF ‚ |

# ISO 8859-14: Extended ASCII (latin8)

|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A0 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | AA | AB | AC | AD | AE | AF |
|    | ·B | ·b | £  | ·C | ·c | ·D | ·S | ·W | ©  | ·W | ·d | ·Y | -  | ®  | ·Y |
| B0 | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | BA | BB | BC | BD | BE | BF |
| ·F | ·f | ·G | ·g | ·M | ·m | ¶  | ·P | ·w | ·p | ·w | ·S | ·y | ·W | ·U | ·S |
| C0 | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | CA | CB | CC | CD | CE | CF |
| ·A | ·A | ·A | ·A | ·A | ·A | Æ  | Ç  | È  | É  | Ê  | Ë  | Ì  | Í  | Î  | Ï  |
| D0 | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | DA | DB | DC | DD | DE | DF |
| ·W | ·N | ·O | ·O | ·O | ·O | ·O | ·T | Ø  | ·U | ·U | ·U | ·U | ·Y | ·Y | ·B |
| E0 | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | EA | EB | EC | ED | EE | EF |
| ·a | ·a | ·a | ·a | ·a | ·a | æ  | Ç  | è  | é  | ê  | ë  | ì  | í  | î  | ï  |
| F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | FA | FB | FC | FD | FE | FF |
| ·w | ·ñ | ·O | ·O | ·O | ·O | ·O | ·t | Ø  | ·u | ·U | ·U | ·U | ·y | ·y | ·y |

# ISO 8859-15: Extended ASCII (lating)

|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A0 | A1 | A2 | A3 | A4 | A5 | A6 | A7 | A8 | A9 | AA | AB | AC | AD | AE | AF |
|    | ı  | ϕ  | £  | €  | ¥  | Š  | š  | Š  | ©  | ı̂ | «  | ¬  | -  | ®  | -  |
| B0 | B1 | B2 | B3 | B4 | B5 | B6 | B7 | B8 | B9 | BA | BB | BC | BD | BE | BF |
| °  | ±  | ²  | ³  | Ž  | ı̂ | ı̇ | ·  | ž  | ı  | ı̂ | »  | œ  | œ  | ÿ  | ı̇ |
| C0 | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | CA | CB | CC | CD | CE | CF |
| À  | Á  | Â  | Ã  | Ä  | Å  | Æ  | Ç  | È  | É  | Ê  | Ë  | Ì  | Í  | Î  | Ï  |
| D0 | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 | DA | DB | DC | DD | DE | DF |
| Ð  | Ñ  | Ò  | Ó  | Ô  | Õ  | Ö  | ×  | Ø  | Ù  | Ú  | Û  | Ü  | Ý  | Þ  | ß  |
| E0 | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | EA | EB | EC | ED | EE | EF |
| à  | á  | â  | ã  | ä  | å  | æ  | ç  | è  | é  | ê  | ë  | ì  | í  | î  | ï  |
| F0 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 | F9 | FA | FB | FC | FD | FE | FF |
| ò  | ñ  | õ  | ö  | ô  | ö  | ö  | ÷  | ø  | ù  | ú  | û  | ü  | ý  | þ  | ÿ  |

# Historical PCs (around 1980)

- Graphics card had a bitmap (say  $9 \times 14$  bits) for each character
- Memory contained array with say  $80 \times 25$  or  $80 \times 50$  bytes for the text on the screen
- Graphics card would build video signal by indexing into the 256 bitmaps
- Firmware included initial bitmap used on system startup
- Different manufacturers used different mappings of numbers to characters

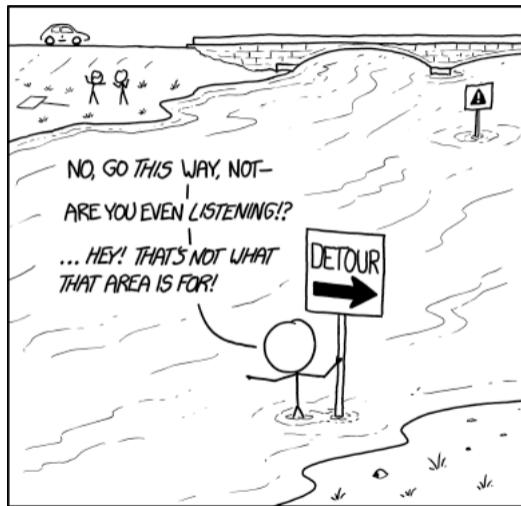
# Major manufacturers create character encodings for major regions (1960-1991)

| V · T · E                             | Character encodings  | [hide] |
|---------------------------------------|--|--------|
| <b>Early telecommunications</b>       | ASCII · ISO/IEC 646 · ISO/IEC 6937 · T.61 · BCDIC · Baudot code · Morse code (Telegraph code · Wabun code) · Special telegraphy codes (Non-Latin · Chinese · Cyrillic) · Needle telegraph codes  |        |
| <b>ISO/IEC 8859</b>                   | -1 · -2 · -3 · -4 · -5 · -6 · -7 · -8 · -9 · -10 · -11 · -12 · -13 · -14 · -15 · -16   |        |
| <b>Bibliographic use</b>              | ANSEL · ISO 5426 / 5426-2 / 5427 / 5428 / 6438 / 6861 / 6862 / 10585 / 10586 / 10754 / 11822 · MARC-8  |        |
| <b>National standards</b>             | ArmSCII · BrasSCII · CNS 11643 · ELOT 927 · GOST 10859 · GB 18030 · HKSCS · ISCII · JIS X 0201 · JIS X 0208 · JIS X 0212 · JIS X 0213 · KOI-7 · KPS 9566 · KS X 1001 · PASCII · SI 960 · TIS-620 · TSCII · VSCII · VSCII · YUSCII  |        |
| <b>EUC</b>                            | CN · JP · KR · TW  |        |
| <b>ISO/IEC 2022</b>                   | CN · JP · KR · CCCII   |        |
| <b>MacOS code pages ("scripts")</b>   | Armenian · Arabic · Barents Cyrillic · Celtic · CentEuro · ChineseSimp / EUC-CN · ChineseTrad / Big5 · Croatian · Cyrillic · Devanagari · Dingbats · Esperanto · Farsi (Persian) · Gaelic · Georgian · Greek · Gujarati · Gurmukhi · Hebrew · Icelandic · Inuit · Japanese / ShiftJIS · Keyboard · Korean / EUC-KR · Latin-1 · Ogham · Roman · Romanian · Sámi · Symbol · Thai / TIS-620 · Turkish · <b>Turkic Latin</b> · Turkic Cyrillic · Ukrainian   |        |
| <b>DOS code pages</b>                 | 100 · 111 · 112 · 113 · 151 · 152 · 161 · 162 · 163 · 164 · 165 · 166 · 210 · 220 · 301 · 437 · 449 · 489 · 620 · 667 · 668 · 707 · 708 · 709 · 710 · 711 · 714 · 715 · 720 · 721 · 737 · 768 · 770 · 771 · 772 · 773 · 774 · 775 · 776 · 777 · 778 · 790 · 850 · 851 · 852 · 853 · 854 · 855/872 · 856 · 857 · 858 · 859 · 860 · 861 · 862 · 863 · 864/17248 · 865 · 866/808 · 867 · 868 · 869 · 874/1161/1162 · 876 · 877 · 878 · 881 · 882 · 883 · 884 · 885 · 891 · 895 · 896 · 897 · 898 · 899 · 900 · 903 · 904 · 906 · 907 · 909 · 910 · 911 · 926 · 927 · 928 · 929 · 932 · 934 · 936 · 938 · 941 · 942 · 943 · 944 · 946 · 947 · 948 · 949 · 950/1370 · 951 · 966 · 991 · 1034 · 1039 · 1040 · 1041 · 1042 · 1043 · 1044 · 1046 · 1086 · 1088 · 1092 · 1093 · 1098 · 1108 · 1109 · 1114 · 1115 · 1116 · 1117 · 1118 · 1119 · 1125/848 · 1126 · 1127 · 1131/849 · 1139 · 1167 · 1168 · 1300 · 1351 · 1361 · 1362 · 1363 · 1372 · 1373 · 1374 · 1375 · 1380 · 1381 · 1385 · 1386 · 1391 · 1393 · 1394 · 1396 · Kamenicky · KOI8 · Mazovia · MIK   |        |
| <b>IBM AIX code pages</b>             | 367 · 371 · 806 · 813 · 819 · 895 · 896 · 912 · 913 · 914 · 915 · 916 · 919 · 920 · 921/901 · 922/902 · 923 · 952 · 953 · 954 · 955 · 956 · 957 · 958 · 959 · 960 · 961 · 963 · 964 · 965 · 970 · 971 · 1004 · 1006 · 1008 · 1009 · 1010 · 1011 · 1012 · 1013 · 1014 · 1015 · 1016 · 1017 · 1018 · 1019 · 1029 · 1036 · 1089 · 1111 · 1124 · 1129/1163 · 1133 · 1350 · 1382 · 1383   |        |
| <b>IBM Apple Macintosh emulations</b> | 1275 · 1280 · 1281 · 1282 · 1283 · 1284 · 1285 · 1286  |        |
| <b>IBM Adobe emulations</b>           | 1038 · 1276 · 1277   |        |
| <b>IBM DEC emulations</b>             | 1020 · 1021 · 1023 · 1090 · 1100 · 1101 · 1102 · 1103 · 1104 · 1105 · 1106 · 1107 · 1287 · 1288  |        |
| <b>IBM HP emulations</b>              | 1050 · 1051 · 1052 · 1053 · 1054 · 1055 · 1056 · 1057 · 1058   |        |
| <b>Windows code pages</b>             | CER-GS · 874/1162 (TIS-620) · 932/943 (Shift JIS) · 936/1386 (GBK) · 950/1370 (Big5) · 949/1363 (EUC-KR) · 1169 · 1174 · Extended Latin-8 · 1200 (UTF-16LE) · 1201 (UTF-16BE) · 1250 · 1251 · 1252 · 1253 · 1254 · 1255 · 1256 · 1257 · 1258 · 1259 · 1261 · 1270 · 54936 (GB18030)  |        |
| <b>EBCDIC code pages</b>              | 1 · 2 · 3 · 4 · 5 · 6 · 7 · 8 · 9 · 10 · 11 · 12 · 13 · 14 · 15 · 16 · 17 · 18 · 19 · 20 · 21 · 22 · 23 · 24 · 25 · 26 · 27 · 28 · 29 · 30 · 31 · 32 · 33 · 34 · 35 · 36 · 37/1140 · 37-2 · 38 · 39 · 40 · 251 · 252 · 254 · 256 · 257 · 258 · 259 · 260 · 264 · 273/1141 · 274 · 275 · 276 · 277/1142 · 278/1143 · 279 · 280/1144 · 281 · 282 · 283 · 284/1145 · 285/1146 · 286 · 287 · 288 · 289 · 290 · 293 · 297/1147 · 298 · 300 · 310 · 320 · 321 · 322 · 330 · 351 · 352 · 353 · 355 · 357 · 358 · 359 · 360 · 361 · 363 · 382 · 383 · 384 · 385 · 386 · 387 · 388 · 389 · 390 · 391 · 392 · 393 · 394 · 395 · 410 · 420/16804 · 421 · 423 · 424/8616/12712 · 425 · 435 · 500/1148 · 803 · 829 · 833 · 834 · 835 · 836 · 837 · 838/838 · 839 · 870/1110/1153 · 871/1149 · 875/4971/9067 · 880 · 881 · 882 · 883 · 884 · 885 · 886 · 887 · 888 · 889 · 890 · 892 · 893 · 905 · 918 · 924 · 930/1390 · 931 · 933/1364 · 935/1388 · 937/1371 · 939/1399 · 1001 · 1002 · 1003 · 1005 · 1007 · 1024 · 1025/1154 · 1026/1155 · 1027 · 1028 · 1030 · 1031 · 1032 · 1033 · 1037 · 1047 · 1068 · 1069 · 1070 · 1071 · 1073 · 1074 · 1075 · 1076 · 1077 · 1078 · 1079 · 1080 · 1081 · 1082 · 1083 · 1084 · 1085 · 1087 · 1091 · 1097 · 1112/1156 · 1113 · 1122/1157 · 1123/1158 · 1130/1164 · 1132 · 1136 · 1137 · 1150 · 1151 · 1152 · 1159 · 1165 · 1166 · 1278 · 1279 · 1303 · 1364 · 1376 · 1377 · JEF · KEIS |        |
| <b>Platform specific</b>              | Acorn · Adobe Standard · Adobe Latin 1 · Apple II · ATASCII · Atari ST · BICS · Casio calculators · CDC · CPC · DEC Radix-50 · DEC MCS/NRCS · DG International · ELWRO-Junior · FIELDATA · GEM · GEOS · GSM 03.38 · HP Roman Extension · HP Roman-8 · HP Roman-9 · HP FOCAL · HP RPL · LICS · LMBCS · Mattel Aquarius · MSX · NEC APC · NeXT · PCW · PETSII · Sharp calculators · TI calculators · TRS-80 · Ventura International · Ventura Symbol · WISCII · XCCS · ZX80 · ZX81 · ZX Spectrum   |        |
| <b>Unicode / ISO/IEC 10646</b>        | UTF-1 · UTF-7 · UTF-8 · UTF-16 (UTF-16LE/UTF-16BE) / UCS-2 · UTF-32 (UTF-32LE/UTF-32BE) / UCS-4 · UTF-EBCDIC · GB 18030 · BOCU-1 · CESU-8 · SCSU   |        |
| <b>Miscellaneous code pages</b>       | ARIB CP 101 · ARIB STD B24 · CoK · HZ · JIS X 0202 · JIS X 0208 · JIS X 0212 · JIS X 0213 · ISO 8859 · ISO 8859-1 · ISO 8859-2 · ISO 8859-3 · ISO 8859-4 · ISO 8859-5 · ISO 8859-6 · ISO 8859-7 · ISO 8859-8 · ISO 8859-9 · ISO 8859-10 · ISO 8859-11 · ISO 8859-12 · ISO 8859-13 · ISO 8859-14 · ISO 8859-15 · ISO 8859-16 · ISO 8859-17 · ISO 8859-18 · ISO 8859-19 · ISO 8859-20 · ISO 8859-21 · Johab · LGR · LY1 · OML · OMS · OMX · OT1 · OT2 · OT3 · OT4 · T2A · T2B · T2C · T2D · T3 · T4 · T5 · T51 · T53 · U · X2 · SEASCII · TACEL8 · TRON · UTF-3 · UTF-6 · WTF-8  |        |
| <b>Related topics</b>                 | Code page · Control character (C0 C1) · CCSID · Character encodings in HTML · Charset detection · Han unification · Hardware · ISO 6429/IEC 6429/ANSI X3.64 · Mojibake   |        |

# Unicode

# Unicode

# Mission impossible



WATCHING THE UNICODE PEOPLE TRY TO GOVERN THE  
INFINITE CHAOS OF HUMAN LANGUAGE WITH CONSISTENT  
TECHNICAL STANDARDS IS LIKE WATCHING HIGHWAY

# 1991: We urgently need a standard for characters...

- ISO 10646 project
- US consortium: Unicode project

## 1991: We urgently need a standard for characters...

- ISO 10646 project
- US consortium: Unicode project

Fortunately:

Unicode 1.1 = ISO 10646-1:1993

Unicode 3.0 = ISO 10646-1:2000

Unicode 7.0 = ISO 10646:2014

... and all are backwards-compatible since Unicode 2.0.

# Unicode uses a 5 layer architecture

1. Abstract Character Repertoire
2. Coded Character Set
3. Character Encoding Form
4. Character Encoding Scheme
5. Transfer Encoding Syntax

# Abstract character repertoire (ACR)

An ACR specifies:

- set of unique characters
- each character has a general name (“at sign”)
- each character has a graphical representation (“@”)

It leaves open:

- representation of characters
- order of characters

# Coded character set (CCS)

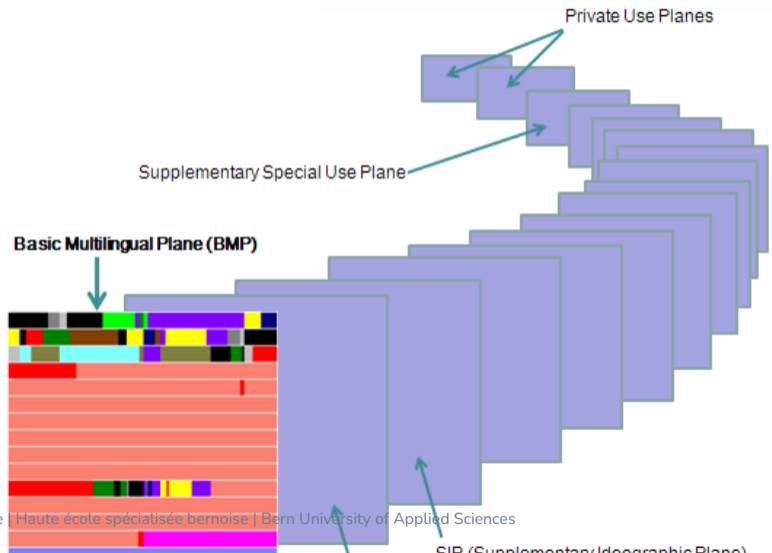
- 1:1 mapping of ACR to positive numbers
- each coded character has a numeric code
- each coded character has a standardized name (“COMMERCIAL AT”)
- each coded character has a code position

Note that a CCS can have holes with positions left unspecified.

<http://www.unicode.org/charts/>

# Basic Multilingual Plane (BMP)

First 65536 code points of Unicode are called the BMP:



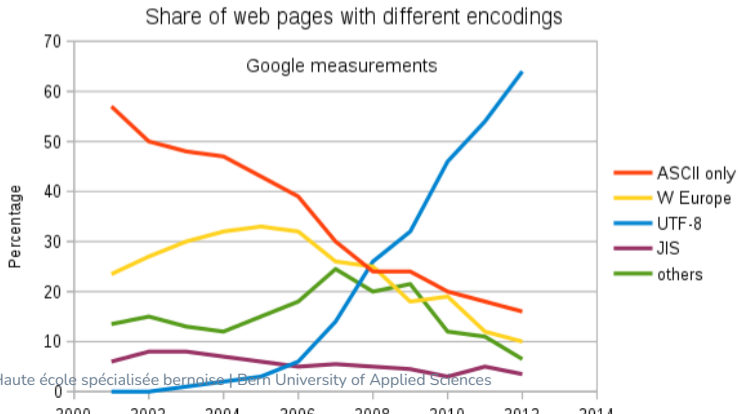
# Character Encoding Form (CEF)

- Specifies mapping of code numbers to code units
- Code units are bit sequences of **fixed length** (usually 8, 16, 24, 32, etc.)

# Character Encoding Scheme (CES)

- Mapping from code units to serialized byte sequences
- Byte order (big endian, little endian) matters
- Examples: UTF-16BE vs UTF-16LE<sup>1</sup>

UTF-8 is the most widely used CES:



# Example

| Name       | Latin A   | Hebrew alef | Han AN      |
|------------|-----------|-------------|-------------|
| Code point | U+0041    | U+05D0      | U+597D      |
| UTF-8      | 41H       | D7H 90H     | E5H A5H BDH |
| UTF-16BE   | 0 41H     | 5H D0H      | 59H 7DH     |
| UTF-32BE   | 0 0 0 41H | 0 0 5H D0H  | 0 0 59H 7DH |

# Transfer Encoding Syntax (TES)

Invertible conversion of byte sequences to:

- eliminate certain undesirable byte sequences that might confuse transfer protocols
- reduce bandwidth consumption via compression (gzip, deflate)

# Example

1. ACR specifies collection of characters, i.e. “a” “!”, “ä” and “%”.
2. CCS specifies numeric codes, i.e. ISO 10646 uses 97, 33, 228 and 8240 for the characters above.
3. A CEF specifies that the above codes are represented using two bytes (UCS-2) or four bytes (UCS-4).
4. A CES may specify how the two bytes are encoded, i.e in big-endian, i.e. for UTF-16BE: 0, 97, 0, 33, 0, 228, 32, 48.
5. A TES may for example apply gzip compression to the above sequence.

# UTF-8

# UTF-8: Motivation

- Most transmitted characters still from 7-bit ASCII
- Common characters in text are in BMP
- Using 32 bits per character would be inefficient!
- C programming language does not deal well with o in strings!

# UTF-8: Rules

- 7-bit ASCII characters use one byte (0xxxxxxx)
- Up to  $2^{11}$  code points use two bytes (110xxxxx.10xxxxxx)
- Up to  $2^{16}$  code points use three bytes (1110xxxx.10xxxxxx.10xxxxxx)
- Up to  $2^{21}$  code points use four bytes (11110xxx.10xxxxxx.10xxxxxx.10xxxxxx)
- Etc.

Note that UTF-8 does not require the shortest possible representation!

oxFE and oxFF cannot appear in UTF-8.  
ox00 only appears for o.

# Conclusion

# Conclusion

- **UTF-8 is the standard now**
  - Used in Web sites
- **Problem for Databases**
  - UTF-8 characters do not have the same length
  - Not possible to be stored in a DB
  - MySQL / MariaDB : Characters are encoded on 3 bytes.
  - MS SQL Server found another solution.
- **Legacy systems use many different encodings**
  - Unix / Mac / Windows
  - US / latin-1 / latin-???
  - International systems : UTF-16BE, UTF-16LE, UTF-32...
- **Job of the Medical Informatics specialist**
  - Let all the systems work together!