# Codifiche di caratteri

**Gli standard principali**

- ASCII (e varianti)
- La famiglia ISO 8859
- ISO/IEC 10646, UCS e Unicode

**Ma anche glifi e fonts**

# Tre concetti indipendenti ma interconnessi

€ Il repertorio di caratteri

£ Il codice

 Le codifiche

# Il repertorio di caratteri

**Semplicemente un insieme di caratteri.**

**Non è necessariamente un alfabeto, ma un alfabeto è un buon esempio:**

- **greco**
- **latino**
- **cirillico**

# I codici

**I codici codificano i caratteri dunque mettono in corrispondenza biunivoca i caratteri con dei numeri.**

**A ciascun elemento (carattere) è associato**
**Il suo codice numerico detto "code position"**

# La codifica

La codifica è la mappatura tra code position e bytes e dipende dai casi

- sotto i 256 simboli viene fatta con un byte
- sotto i 65536 simboli con due byte
- oltre con altri byte o con codifiche complesse

# Codifica ASCII

**ASCII: American Standard Code for Information Interchange**

**È uno standard ANSI che mappa valori per 128 caratteri usando 7 bit su 8 di un byte (il primo bit - a sinistra non è significativo, viene usato per il controllo di parità.**

**Il termine ASCII è utilizzato sia per i caratteri che per la codifica**

# ASCII

**33 caratteri (0-31 e 127) sono di controllo, con ripetizioni**

Backspace (sposta la testina indietro di un carattere, utile nelle telescriventi - 08 [0x08]) e

Delete (cancella tutti i buchi di un carattere in una scheda perforata, cioè tutti buchi, 1111111 - 127 [0x7F]).

Carriage Return (riporta la testina all'inizio di riga - 13 [0x0C]) e

Form Feed (gira il carrello di una riga - 14 [0x0D])

# ASCII

Gli altri 95, composti da caratteri dell'alfabeto latino, maiuscole e minuscole, numeri e punteggiatura sono qui di seguito rappresentati (il primo e l'ultimo sono spazi bianchi):

```
 !"#$%&'()*+,-./
0123456789:;<=>?
@ABCDEFGHIJKLMNO
PQRSTUVWXYZ[\]^_
`abcdefghijklmno
pqrstuvwxyz{|}~
```

# Variazioni dell'ASCII

Esistono molte variazioni internazionali dell'ASCII.

In queste variazioni alcuni caratteri sono sostituiti da
• simboli speciali propri della nazione che ha creato la variante.

Qualche volta si parla dell'ASCII a "8 bit" questo nome è utilizzato per indicare varie codifiche che sono estensioni dell'ASCII nel senso che contengono ASCII come sottoinsieme ma utilizzano l'intervallo 128-255 per estendere l'insieme.

| Byte | Cod. | Char |
|---|---|---|
| 00000000 | 0 | Null |
| 00000001 | 1 | Start of heading |
| 00000010 | 2 | Start of text |
| 00000011 | 3 | End of text |
| 00000100 | 4 | End of transmit |
| 00000101 | 5 | Enquiry |
| 00000110 | 6 | Acknowledge |
| 00000111 | 7 | Audible bell |
| 00001000 | 8 | Backspace |
| 00001001 | 9 | Horizontal tab |
| 00001010 | 10 | Line feed |
| 00001011 | 11 | Vertical tab |
| 00001100 | 12 | Form Feed |
| 00001101 | 13 | Carriage return |
| 00001110 | 14 | Shift out |
| 00001111 | 15 | Shift in |
| 00010000 | 16 | Data link escape |
| 00010001 | 17 | Device control 1 |
| 00010010 | 18 | Device control 2 |
| 00010011 | 19 | Device control 3 |
| 00010100 | 20 | Device control 4 |
| 00010101 | 21 | Neg. acknowledge |
| 00010110 | 22 | Synchronous idle |
| 00010111 | 23 | End trans. block |
| 00011000 | 24 | Cancel |
| 00011001 | 25 | End of medium |
| 00011010 | 26 | Substitution |
| 00011011 | 27 | Escape |
| 00011100 | 28 | File separator |
| 00011101 | 29 | Group separator |
| 00011110 | 30 | Record Separator |
| 00011111 | 31 | Unit separator |

| Byte | Cod. | Char |
|---|---|---|
| 00100000 | 32 | Spc |
| 00100001 | 33 | ! |
| 00100010 | 34 | " |
| 00100011 | 35 | # |
| 00100100 | 36 | $ |
| 00100101 | 37 | % |
| 00100110 | 38 | & |
| 00100111 | 39 | ' |
| 00101000 | 40 | ( |
| 00101001 | 41 | ) |
| 00101010 | 42 | * |
| 00101011 | 43 | + |
| 00101100 | 44 | , |
| 00101101 | 45 | - |
| 00101110 | 46 | . |
| 00101111 | 47 | / |
| 00110000 | 48 | 0 |
| 00110001 | 49 | 1 |
| 00110010 | 50 | 2 |
| 00110011 | 51 | 3 |
| 00110100 | 52 | 4 |
| 00110101 | 53 | 5 |
| 00110110 | 54 | 6 |
| 00110111 | 55 | 7 |
| 00111000 | 56 | 8 |
| 00111001 | 57 | 9 |
| 00111010 | 58 | : |
| 00111011 | 59 | ; |
| 00111100 | 60 | < |
| 00111101 | 61 | = |
| 00111110 | 62 | > |
| 00111111 | 63 | ? |

| Byte | Cod. | Char |
|---|---|---|
| 01000000 | 64 | @ |
| 01000001 | 65 | A |
| 01000010 | 66 | B |
| 01000011 | 67 | C |
| 01000100 | 68 | D |
| 01000101 | 69 | E |
| 01000110 | 70 | F |
| 01000111 | 71 | G |
| 01001000 | 72 | H |
| 01001001 | 73 | I |
| 01001010 | 74 | J |
| 01001011 | 75 | K |
| 01001100 | 76 | L |
| 01001101 | 77 | M |
| 01001110 | 78 | N |
| 01001111 | 79 | O |
| 01010000 | 80 | P |
| 01010001 | 81 | Q |
| 01010010 | 82 | R |
| 01010011 | 83 | S |
| 01010100 | 84 | T |
| 01010101 | 85 | U |
| 01010110 | 86 | V |
| 01010111 | 87 | W |
| 01011000 | 88 | X |
| 01011001 | 89 | Y |
| 01011010 | 90 | Z |
| 01011011 | 91 | [ |
| 01011100 | 92 | \ |
| 01011101 | 93 | ] |
| 01011110 | 94 | ^ |
| 01011111 | 95 | _ |

| Byte | Cod. | Char |
|---|---|---|
| 01100000 | 96 | ` |
| 01100001 | 97 | a |
| 01100010 | 98 | b |
| 01100011 | 99 | c |
| 01100100 | 100 | d |
| 01100101 | 101 | e |
| 01100110 | 102 | f |
| 01100111 | 103 | g |
| 01101000 | 104 | h |
| 01101001 | 105 | i |
| 01101010 | 106 | j |
| 01101011 | 107 | k |
| 01101100 | 108 | l |
| 01101101 | 109 | m |
| 01101110 | 110 | n |
| 01101111 | 111 | o |
| 01110000 | 112 | p |
| 01110001 | 113 | q |
| 01110010 | 114 | r |
| 01110011 | 115 | s |
| 01110100 | 116 | t |
| 01110101 | 117 | u |
| 01110110 | 118 | v |
| 01110111 | 119 | w |
| 01111000 | 120 | x |
| 01111001 | 121 | y |
| 01111010 | 122 | z |
| 01111011 | 123 | { |
| 01111100 | 124 | \| |
| 01111101 | 125 | } |
| 01111110 | 126 | ~ |
| 01111111 | 127 | Del |

# Tabella ASCII Estesa

| Byte | Cod. | Char |
|---|---|---|
| 10000000 | 128 | Ç |
| 10000001 | 129 | ü |
| 10000010 | 130 | é |
| 10000011 | 131 | â |
| 10000100 | 132 | ä |
| 10000101 | 133 | à |
| 10000110 | 134 | å |
| 10000111 | 135 | ç |
| 10001000 | 136 | ê |
| 10001001 | 137 | ë |
| 10001010 | 138 | è |
| 10001011 | 139 | ï |
| 10001100 | 140 | î |
| 10001101 | 141 | ì |
| 10001110 | 142 | Ä |
| 10001111 | 143 | Å |
| 10010000 | 144 | É |
| 10010001 | 145 | æ |
| 10010010 | 146 | Æ |
| 10010011 | 147 | ô |
| 10010100 | 148 | ö |
| 10010101 | 149 | ò |
| 10010110 | 150 | û |
| 10010111 | 151 | ù |
| 10011000 | 152 | ÿ |
| 10011001 | 153 | Ö |
| 10011010 | 154 | Ü |
| 10011011 | 155 | ø |
| 10011100 | 156 | £ |
| 10011101 | 157 | Ø |
| 10011110 | 158 | × |
| 10011111 | 159 | ƒ |

| Byte | Cod. | Char |
|---|---|---|
| 10100000 | 160 | á |
| 10100001 | 161 | í |
| 10100010 | 162 | ó |
| 10100011 | 163 | ú |
| 10100100 | 164 | ñ |
| 10100101 | 165 | Ñ |
| 10100110 | 166 | ª |
| 10100111 | 167 | º |
| 10101000 | 168 | ¿ |
| 10101001 | 169 | ® |
| 10101010 | 170 | ¬ |
| 10101011 | 171 | ½ |
| 10101100 | 172 | ¼ |
| 10101101 | 173 | ¡ |
| 10101110 | 174 | « |
| 10101111 | 175 | » |
| 10110000 | 176 | ▒ |
| 10110001 | 177 | ▒ |
| 10110010 | 178 | ▒ |
| 10110011 | 179 | ┤ |
| 10110100 | 180 | ┤ |
| 10110101 | 181 | Á |
| 10110110 | 182 | Â |
| 10110111 | 183 | À |
| 10111000 | 184 | © |
| 10111001 | 185 | ╣ |
| 10111010 | 186 | ║ |
| 10111011 | 187 | ╗ |
| 10111100 | 188 | ╝ |
| 10111101 | 189 | ¢ |
| 10111110 | 190 | ¥ |
| 10111111 | 191 | ┐ |

| Byte | Cod. | Char |
|---|---|---|
| 11000000 | 192 | └ |
| 11000001 | 193 | ┴ |
| 11000010 | 194 | ┬ |
| 11000011 | 195 | ├ |
| 11000100 | 196 | ─ |
| 11000101 | 197 | ┼ |
| 11000110 | 198 | ã |
| 11000111 | 199 | Ã |
| 11001000 | 200 | ╚ |
| 11001001 | 201 | ╔ |
| 11001010 | 202 | ╩ |
| 11001011 | 203 | ╦ |
| 11001100 | 204 | ╠ |
| 11001101 | 205 | ═ |
| 11001110 | 206 | ╬ |
| 11001111 | 207 | ¤ |
| 11010000 | 208 | ð |
| 11010001 | 209 | Ð |
| 11010010 | 210 | Ê |
| 11010011 | 211 | Ë |
| 11010100 | 212 | È |
| 11010101 | 213 | ı |
| 11010110 | 214 | Í |
| 11010111 | 215 | Î |
| 11011000 | 216 | Ï |
| 11011001 | 217 | ┘ |
| 11011010 | 218 | ┌ |
| 11011011 | 219 | █ |
| 11011100 | 220 | ▄ |
| 11011101 | 221 | ¦ |
| 11011110 | 222 | Ì |
| 11011111 | 223 | ▀ |

| Byte | Cod. | Char |
|---|---|---|
| 11100000 | 224 | Ó |
| 11100001 | 225 | ß |
| 11100010 | 226 | Ô |
| 11100011 | 227 | Ò |
| 11100100 | 228 | õ |
| 11100101 | 229 | Õ |
| 11100110 | 230 | µ |
| 11100111 | 231 | þ |
| 11101000 | 232 | Þ |
| 11101001 | 233 | Ú |
| 11101010 | 234 | Û |
| 11101011 | 235 | Ù |
| 11101100 | 236 | ý |
| 11101101 | 237 | Ý |
| 11101110 | 238 | ¯ |
| 11101111 | 239 | ´ |
| 11110000 | 240 |  |
| 11110001 | 241 | ± |
| 11110010 | 242 | ‗ |
| 11110011 | 243 | ¾ |
| 11110100 | 244 | ¶ |
| 11110101 | 245 | § |
| 11110110 | 246 | ÷ |
| 11110111 | 247 | ¸ |
| 11111000 | 248 | ° |
| 11111001 | 249 | ¨ |
| 11111010 | 250 | · |
| 11111011 | 251 | ¹ |
| 11111100 | 252 | ³ |
| 11111101 | 253 | ² |
| 11111110 | 254 | ■ |
| 11111111 | 255 | |

# ISO Latin 1

Una estensione dell'ASCII standard che comprende alcuni caratteri un certo numero di caratteri degli alfabeti europei come accenti, ecc. è l'ISO 8859-1 (della famiglia ISO 8859) che comprende il repertorio di caratteri "Latin alphabet No. 1", noto come ISO Latin 1.

ISO Latin 1 è compatibile all'indietro con ASCII, di cui è un estensione per i soli caratteri >127.

# ISO Latin 1

**I caratteri dell'ISO Latin 1 che occupano le posizioni da 160 a 255:**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A0 | A1 ¡ | A2 ¢ | A3 £ | A4 ¤ | A5 ¥ | A6 ¦ | A7 § | A8 ¨ | A9 © | AA ª | 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 ð | 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 |

# La famiglia ISO 8859:
# Estende ascii per altri alfabeti

| | |
|---|---|
| iso-8859-1, Latin1, L1 | Europa dell'Ovest |
| iso-8859-2, latin2, L2 | Europa centrale e dell'Est |
| iso-8859-3, latin3, L3 | Esperanto, Maltese |
| iso-8859-4, latin4, L4 | Lingue baltiche |
| iso-8859-5, cyrillic | Bulgaro, Bielorusso, Macedone, Russo, Serbo |
| iso-8859-6, arabic | Arabo |
| iso-8859-7, greek, greek8 | Greco |
| iso-8859-8, hebrew | Ebraico |
| iso-8859-9, latin5, L5 | Turco |
| iso-8859-10, latin6, L6 | Lingue nordiche |
| iso-8859-13 | Lingue baltiche |
| iso-8859-14, latin8, L8 | Lingue Celte |
| iso-8859-15 | Europa dell'Ovest (Simbolo dell'Euro €) |

# ISO/IEC 10646, UCS e Unicode

Negli anni '90 sono state avviate due iniziative parallele per sistemare in modo definitivo il problema della rappresentazione dei caratteri:

la prima gestita dal consorzio Unicode (una organizzazione no-profit in cui convergono numerosi produttori di sistemi informatici), e la seconda dall'ISO.

Da queste iniziative sono nati dal consorzio Unicode e ISO 10646/UCS, due codifiche di caratteri che sono perfettamente allineati (le differenze riguardano solo aspetti tecnici) e che possono definirsi 'universali'. UCS sta infatti per "Universal Character Set" e corrisponde ad un vastissimo repertorio di caratteri e il suo corrispondente insieme di codici.

# Unicode

Lo standard Unicode definito dallo Unicode Consortium era originalmente progettato per essere un codice a 16 bit (che equivale a $2^{16}$=65536 caratteri), ma fu esteso in modo da permettere codici nell'intervallo esadecimale 0..10FFFF vale a dire 1 114 112 caratteri.

Tipicamente un carattere Unicode viene identificato con l'abbreviazione U+xxxx dove xxxx è un numero esadecimale a quattro cifre. Ad esempio la lettera *A* viene indicata in Unicode come U+0041 (65 in decimale).

# UCS-2 e UTF-16

**UCS-2 e UTF-16 sono i nomi due codifiche di caratteri quasi identiche.**

**UCS-2 è uno schema a due byte ed è sostanzialmente un'estensione di ISO Latin 1.**

**Ad esempio nel caso della lettera maiuscola *A* rappresentata in ISO Latin 1 con il codice 41 esadecimale in UCS-2 verrebbe rappresentato da 2 bytes 00 41.**

# UCS-2: svantaggi

**L'uso di UCS-2 comporta, però, tre svantaggi:**

- il file occupano il doppio di spazio rispetto a quelli codificati in ISO Latin 1 (due byte invece di uno);
- UCS-2 non è compatibile all'indietro con ASCII (programmi che si aspettano testi codificati con byte singoli non possono leggere testi un UCS-2);
- UCS-2 è limitato a rappresentare massimo 65.536 caratteri.

# UTF-8
# Unicode Transformation Format

UTF-8 è una codifica a lunghezza variabile di Unicode che risolve i primi due problemi dell'UCS-2 di cui si è discusso sopra.

I caratteri da 0 a 127, ovvero il set di caratteri ASCII, vengono codificati con un byte ciascuno, esattamente come avviene nella codifica ASCII.

In ASCII, il byte con valore 65 rappresenta la lettera A e anche in UTF-8 il byte 65 rappresenta la lettera A. Quindi esiste un'identità uno-a-uno tra i caratteri ASCII e i byte di UTF-8.

Questo significa che i file scritti in ASCII puro risultano anche accettabili come file UTF-8.

# UTF-8

**UTF-8 permette di accedere a tutti i caratteri definiti di *UCS-4*, ma utilizza un numero compreso tra 1 e 4 byte per farlo**

- i codici compresi tra 0 - 127 (*ASCII* a 7 bit) richiedono un byte, in cui ci sia 0 al primo bit;
- i codici derivati dall'alfabeto latino e tutti gli script non-ideografici richiedono 2 byte;
- i codici ideografici (orientali) richiedono 3 byte;

# UTF-8

| | |
|---|---|
| U-00000000 – U-0000007F: | 0*xxxxxxx* |
| U-00000080 – U-000007FF: | 110*xxxxx* 10*xxxxxx* |
| U-00000800 – U-0000FFFF: | 1110*xxxx* 10*xxxxxx* 10*xxxxxx* |
| U-00010000 – U-001FFFFF: | 11110*xxx* 10*xxxxxx* 10*xxxxxx* 10*xxxxxx* |

# Glifi e Fonts

**Per mostrarci un testo sullo schermo o stamparlo su carta, il computer ha bisogno di codificare due tipi di informazioni diverse:**

il puro e semplice contenuto del testo (plain text)

la forma o layout del testo, cioè una serie di codici che ne definiscono il formato (numero di colonne, larghezza dei margini, eccetera)
e determinati
attribuiti grafici (il tipo di font, la dimensione del testo, il suo colore, eccetera) - (rich text)

# Glifo

Il glifo è una particolare forma con cui un carattere può essere rappresentato sullo schermo o su carta.

Per esempio, il carattere Z potrebbe essere rappresentato come grassetto **Z** o corsivo *Z*.

D'altra parte, il carattere minuscolo z è definito come un carattere separato, il quale a sua volta potrebbe essere associato ad un glifo diverso.

In Unicode ci sono diversi esempi di caratteri che potrebbero essere considerate semplicemente delle varianti tipografiche di uno stesso carattere, ma che per varie ragioni sono tenuti come caratteri separati.

# Font

Un repertorio di glifi costituisce un font.

Più tecnicamente un font è un insieme numerato di glifi. Il numero corrisponde al codice del carattere (rappresentato dal glifo).

Quindi in certo senso il font è dipendente dal codice del carattere.

È possibile che un glifo utilizzato per un dato carattere venga utilizzate per altri caratteri.