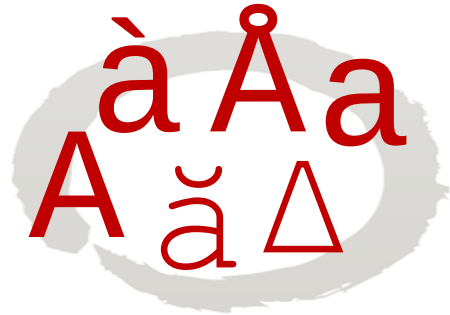
# Testo

lezione quattro

# Testo

- Tra i vari media il **testo** è quello che ha una rappresentazione digitale naturale, essendo nativamente una sequenza di simboli
  - È comunque necessario un sistema di **codifica** che stabilisca come rappresentare i diversi simboli che compongono il testo
  - Inoltre, ogni simbolo può essere rappresentato visivamente in modi diversi, con diversi stili **tipografici**, dimensioni e colori.



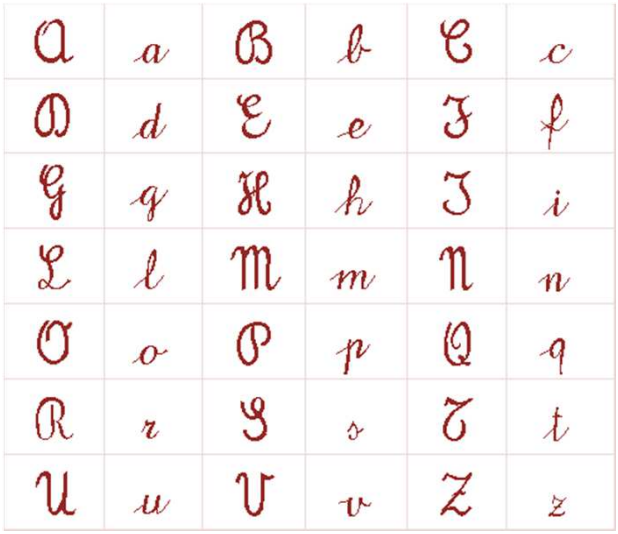

# Testo

- Esamineremo il testo nella sua doppia natura:
  - Prima vedremo i principali aspetti legati alla **tipografia**: il testo ha proprietà grafiche (tra cui per esempio il font, la dimensione, il colore) che determinano come viene visualizzato o stampato. Anche queste caratteristiche devono essere (in alcuni casi come per esempio il WWW) stabili rispetto a protocolli e architettura di Internet.

  - Poi introdurremo gli aspetti legati alla **codifica**: la codifica deve garantire portabilità del contenuto testuale attraverso protocolli e architetture di Internet. Il testo scritto su una specifica architettura deve essere leggibile su una architettura differente.

# Carattere, alfabeto, charset

- **carattere** è un'unità di informazione che, corrisponde un simbolo della forma scritta di una lingua naturale.
- L'insieme di caratteri che vengono usati in una lingua è un **alfabeto**, ovvero un sistema di scrittura i cui segni grafici rappresentano i suoni di una lingua.
- Un insieme di caratteri codificati è invece un **charset**.



# Glifo e fonte

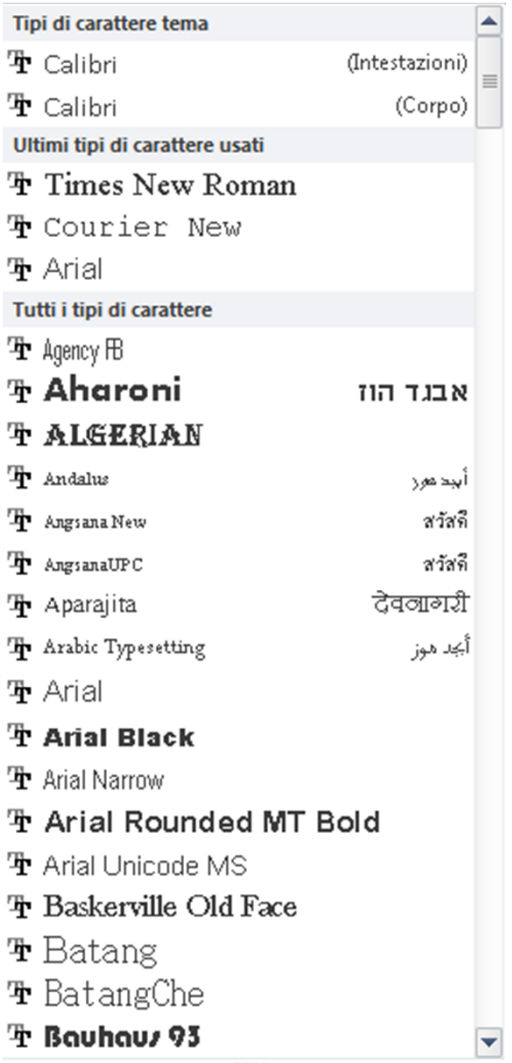
- Il **glifo** è una entità tipografica che realizza la rappresentazione visiva della forma del carattere:
  - Ogni carattere può essere rappresentato da molti glifi differenti (per esempio queste sono tutte P maiuscole: P 𝐏 𝓟 P ℙ)
  - Lo stesso glifo potrebbe rappresentare due caratteri (per esempio il glifo P in cirillico Russo è traslitterato R negli alfabeti latini). •
- Un insieme di glifi che rappresentano i caratteri di un alfabeto (o di un charset) è detta **fonte**

# Legatura

- Inoltre, due caratteri possono corrispondere ad un solo glifo, quando esiste una legatura.
- La **legatura** è l'unione di due o più caratteri che vengono fusi in un'unica forma grafica.
- Le legature di solito rimpiazzano due caratteri che condividono uno spazio comune
- Un esempio di legatura è fi → fi

# Font

- Un **font** è quindi un insieme di glifi caratterizzati da un certo stile grafico o progettati per svolgere una data funzione:
  - Ogni font contiene un certo numero di glifi, che rappresentano lettere, numeri e punteggiatura.
  - In sostanza un font è, in informatica, una rappresentazione grafica di un insieme di caratteri, cioè di un charset



# Font digitali

- Ci sono 3 tipologie di font digitali:
  - Font **bitmap**: ogni glifo è realizzato da una matrice di punti.
  - Font **vettoriali** (o **outline** font): ogni glifo è definito attraverso curve di Bézier (vettori) che ne delineano i contorni. I font true type vanno rasterizzati per essere riprodotti a schermo.
  - Font **stroke**: ogni glifo è definito dai vertici di tratti individuali. Si riduce il numero di vertici necessaria a riprodurre un glifo e si aumenta la scalabilità.

# Font vettoriali

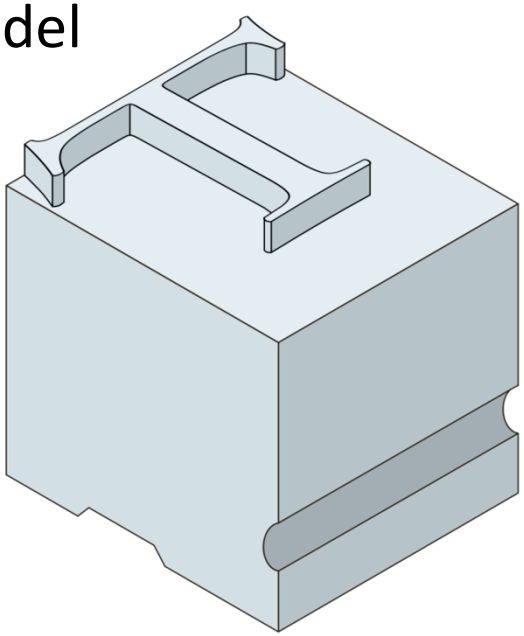
- Tra più importanti font vettoriali ci sono:
  - **Type 1**: formato di caratteri ideato da Adobe e incluso in Postscript. Prevede un meccanismo di rasterizzazione proprietario.
  - **Type 3**: sostanzialmente equivalente a Type 1 ma senza rasterizzazione proprietaria (Adobe)

  - **Truetype**: proposto da Apple e adottato anche da Windows, sviluppato in competizione con Type 1

  - **Opentype**: sviluppato da Microsoft e da Adobe (ma pensato per essere completamente cross platform) ha caratteristiche tipografiche più avanzate dei predecessori:
    - La codifica è basata su Unicode.
    - i font possono avere caratteristiche tipografiche avanzate che consentono la gestione di linguaggi che usano alfabeti non latini.

# Dimensione dei font

- L'unità di misura per le dimensioni del corpo del testo più usata è il **punto tipografico** (**pt**) corrispondente ad 1/72 di pollice.
- Nella tipografia tradizionale, la dimensione del corpo era misurata in base all'altezza del blocchetto di piombo usato per imprimere il carattere sulla carta.



- Per misurare i caratteri si usano:
  - **linea di base**: linea orizzontale immaginaria su cui si appoggiano i caratteri.
  - **occhio medio**, la distanza fra la linea base e la cima di un normale carattere minuscolo

# Dimensione dei font

- La dimensione è misurata così (a causa di effetti ottici che rendono le lettere tonde inadatte):
  - Le minuscole si misurano solitamente sulla lettera x.
  - Le maiuscole si misurano solitamente sulla lettera E
  - Ascendente, parte delle lettere minuscole alte (come l, t, f) che possono essere anche più alte delle maiuscole.
  - Discendente: parte delle lettere minuscole che scende sotto la linea di base
  - Per garantire una minima interlinea vengono lasciate libere la spalla superiore e la spalla inferiore

# Caratteristiche dei font

- I font sono classificabili secondo diverse caratteristiche:
  - **proporzionale/monospace**:
    - se i glifi hanno lunghezza variabile i font sono proporzionali. Es: Arial
    - se hanno larghezza fissa sono monospace (a larghezza fissa). Es. Courier New


  - **serif/sans-serif**:
    - Se i glifi hanno le grazie sono serif. Es: Arial
    - Se non hanno le grazie sono sans-serif. Es: Times New Roman

# Serif e Sans-serif

- Le **grazie** (serif, in inglese) sono allungamenti ortogonali delle estremità di un glifo.
- Le grazie nascono dalla necessità che avevano gli scalpellini dell'antica Roma di incidere le lettere nella pietra, scolpendo terminazioni delle lettere ad angolo retto
- Ci sono diversi tipi di grazie:
  - Bodoni, a bottone: f c r
  - Garamond, a goccia: f c r
  - Palatino, a becco: f c r
- I font con le grazie sono detti **font serif**, quelli senza le grazie, **font sans-serif** (o a volte solo sans).

# Famiglia di font

- **Font family** (famiglia di font) è un insieme di stili diversi di uno stesso carattere.
- **Generic family** è invece un insieme di font family accomunati da caratteristiche simili.
- Le variazioni sono basate su un unico design che solitamente è variato in base a:
  - peso (bold, normal, light, extralight)
  - inclinazione dell'asse (roman, italic)
  - presenza/assenza di grazie.

| Generic family | Font family |
|---|---|
| Serif | Times New Roman<br>Georgia |
| Sans-serif | Arial<br>Verdana |
| Monospace | Courier New<br>Lucida Console |

Arial Regular
**Arial Bold**
*Arial Italic*
***Arial Bold Italic***
Arial Narrow Regular
Arial Narrow Bold
*Arial Narrow Italic*
***Arial Narrow Bold Italic***
**Arial Black Regular**
***Arial Black Italic***

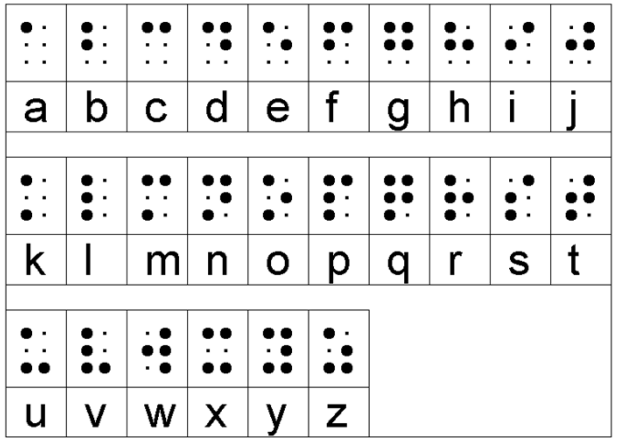
# … cambiamo punto di vista

- Tipografia e tipometria sono argomenti complessi che potrebbero essere approfonditi di più.
  - Li vedrete in pratica nelle lezioni sui CSS.
  - Ci sono vari approfondimenti sul libro o nelle risorse on line.
- Invece ora passiamo a trattare il testo dal punto di vista della **codifica**: poiché consideriamo un contesto di rete ogni carattere deve essere codificato in modo da mantenere le proprie caratteristiche anche a dopo la comunicazione ad un host diverso da quello in cui era originariamente memorizzato.
- Questa proprietà non è affatto scontata.

# Codifica dei caratteri

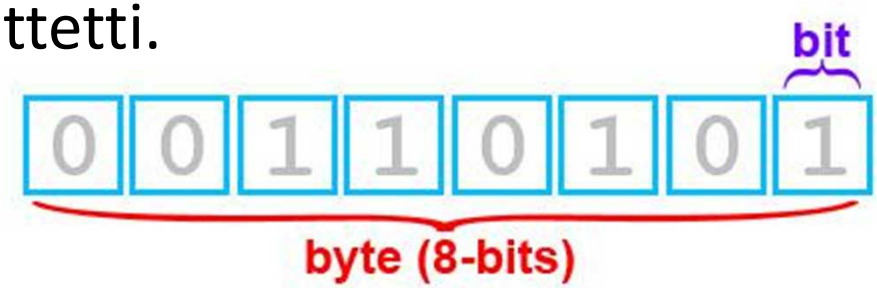
- I caratteri sono codificati attraverso codici che mappano ciascun carattere di un certo insieme di «*codici*»
- I sistemi di codifica dei caratteri esistevano prima dell'avvento dei computer.
- Esempi sono:
  - Codice **Morse**: usato per trasmettere attraverso un segnale ad intermittenza
  - Codice **Braille**: usato per dare rappresentazione tattile ai caratteri

| Lettere | Codice | Lettere | Codice | Numeri | Codice | Punteg. | Codice |
|---|---|---|---|---|---|---|---|
| A | ·— | N | —· | 0 | ————— | · | ·—·—·— |
| B | —··· | O | ——— | 1 | ·———— | , | ——··—— |
| C | —·—· | P | ·——· | 2 | ··——— | : | ———··· |
| D | —·· | Q | ——·— | 3 | ···—— | ? | ··——·· |
| E | · | R | ·—· | 4 | ····— | = | —···— |
| F | ··—· | S | ··· | 5 | ····· | - | —····— |
| G | ——· | T | — | 6 | —···· | ( | —·——· |
| H | ···· | U | ··— | 7 | ——··· | ) | —·——·— |
| I | ·· | V | ···— | 8 | ———·· | " | ·—··—· |
| J | ·——— | W | ·—— | 9 | ————· | ' | ·————· |
| K | —·— | X | —··— | | | / | —··—· |
| L | ·—·· | Y | —·—— | | | Sottolineato | ··——·— |
| M | —— | Z | ——·· | | | @ | ·——·—· |



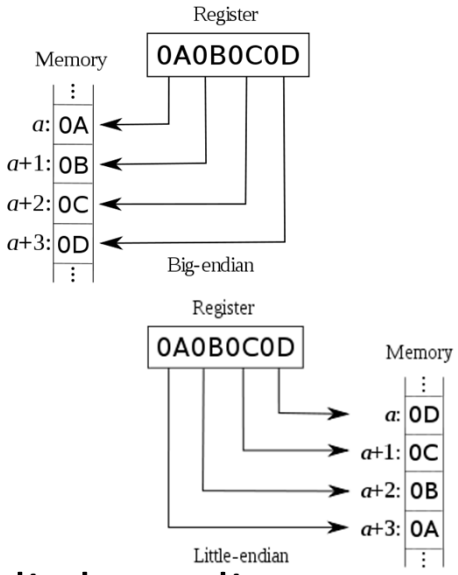
# Codifica dei caratteri

- In informatica i caratteri sono codificati attraverso codici che mappano ciascun carattere di un certo insieme (**charset**) in bit o byte:
  - **bit**, binary digit: uno dei due simboli del sistema numerico binario (0,1); un bit di memoria consente la codifica di 2 stati.
  - **byte** (o meglio ottetti, **octet**) gruppo composto da 8 bit. Un byte di memoria consente la codifica 256 diversi ottetti.



# Codifica

- Da Architetture riprendiamo che ci sono due modi (principali) di ordinare i byte che costituiscono un dato:
  - **big-endian**: memorizzazione che inizia dal byte **più** significativo per finire col meno significativo; è usata, per esempio, dai processori **Motorola**;
  - **little-endian**: memorizzazione che inizia dal byte **meno** significativo per finire col più significativo; è utilizzata, per esempio, dai processori **Intel**.




- In rete si distinguono:
  - **host byte order** (che può essere big-endian o little-endian o altro), l'ordine nativo dell'host
  - **network byte ordern** (big-endian), l'ordine standard in molti protocolli Internet

# Il codice ASCII

- ASCII (American Standard Code for Information Interchange) è una codifica:
  - Definita nel 1961
  - pubblicato dall'American National Standards Institute (ANSI) nel 1968,
  - Diventata standard ISO (ISO 646) nel 1972.
- E' una codifica basata su 7 bit: si usa un byte di memoria ma gli ottetti da 128 a 255 non sono utilizzati.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 000 | nul | 001 | soh | 002 | stx | 003 | etx | 004 | eot | 005 | enq | 006 | ack | 007 | bel |
| 008 | bs | 009 | ht | 010 | nl | 011 | vt | 012 | np | 013 | cr | 014 | so | 015 | si |
| 016 | dl | 017 | dc1 | 018 | dc2 | 019 | dc3 | 020 | dc4 | 021 | nak | 022 | syn | 023 | etb |
| 024 | can | 025 | em | 026 | sub | 027 | esc | 028 | fs | 029 | gs | 030 | rs | 031 | us |
| 032 | sp | 033 | ! | 034 | " | 035 | # | 036 | $ | 037 | % | 038 | & | 039 | ' |
| 040 | ( | 041 | ) | 042 | * | 043 | + | 044 | , | 045 | - | 046 | . | 047 | / |
| 048 | 0 | 049 | 1 | 050 | 2 | 051 | 3 | 052 | 4 | 053 | 5 | 054 | 6 | 055 | 7 |
| 056 | 8 | 057 | 9 | 058 | : | 059 | ; | 060 | < | 061 | = | 062 | > | 063 | ? |
| 064 | @ | 065 | A | 066 | B | 067 | C | 068 | D | 069 | E | 070 | F | 071 | G |
| 072 | H | 073 | I | 074 | J | 075 | K | 076 | L | 077 | M | 078 | N | 079 | O |
| 080 | P | 081 | Q | 082 | R | 083 | S | 084 | T | 085 | U | 086 | V | 087 | W |
| 088 | X | 089 | Y | 090 | Z | 091 | [ | 092 | \ | 093 | ] | 094 | ^ | 095 | _ |
| 096 | ` | 097 | a | 098 | b | 099 | c | 100 | d | 101 | e | 102 | f | 103 | g |
| 104 | h | 105 | i | 106 | j | 107 | k | 108 | l | 109 | m | 110 | n | 111 | o |
| 112 | p | 113 | q | 114 | r | 115 | s | 116 | t | 117 | u | 118 | v | 119 | w |
| 120 | x | 121 | y | 122 | z | 123 | { | 124 | \| | 125 | } | 126 | ~ | 127 | del |

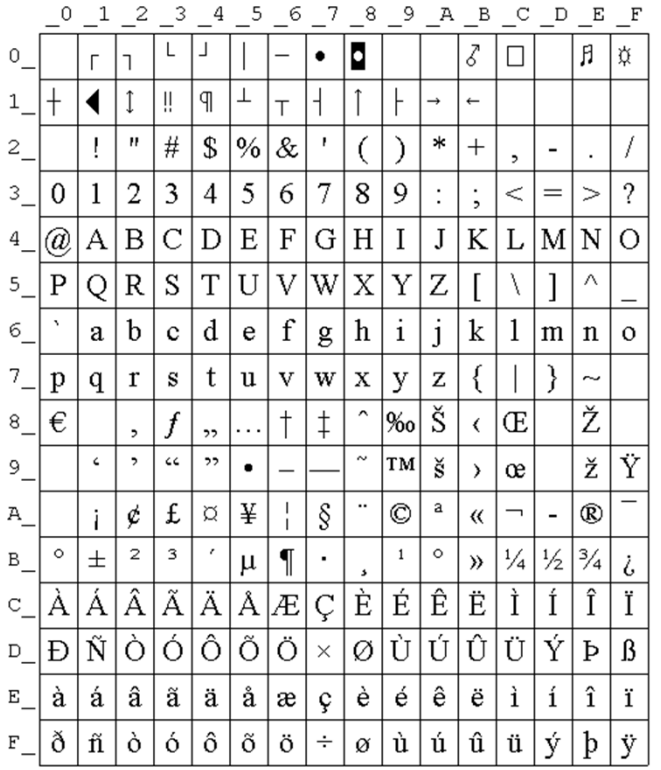
# Varianti nazionali di ISO 646

- Il codice ASCII non contiene alcuni caratteri molto usati in alcune lingue europee (per esempio tutte le lettere accentate).
- In ISO 646 sono definite anche **varianti nazionali**, in cui alcune posizioni sono assegnate per uso nazionale
- Queste posizioni sono:
  - **@[\]{|}** sempre e
  - **#$^`~** se necessario.

| dec | hex | glifo | variante |
|---|---|---|---|
| 35 | 23 | # | £ Ù |
| 36 | 24 | $ | ¤ |
| 64 | 40 | @ | É § Ä à ³ |
| 91 | 5B | [ | Ä Æ ° â ¡ ÿ é |
| 92 | 5C | \ | Ö Ø ç Ñ ½ ¥ |
| 93 | 5D | ] | Å Ü § ê é ¿ \| |
| 94 | 5E | ^ | Ü î è |
| 96 | 60 | ` | é ä µ ô ù |
| 123 | 7B | { | ä æ é à ° ¨ |
| 124 | 7C | \| | ö ø ù ò ñ *f* |
| 125 | 7D | } | å ü è ç ¼ |
| 126 | 7E | ~ | ü ¯ ß ¨ û ì ´ _ |

# ISO 8859/1 (ISO Latin 1)

- **ISO Latin 1** è uno standard che compone ISO 8859, e come tutte le specifiche ISO 8859 utilizza 8 bit
- I primi 128 caratteri sono quelli di ASCII, gli altri 128 sono usati per introdurre i caratteri latini specifici.
- Copre:
  - la maggior parte delle lingue europee occidentali: danese, faroese, finlandese, francese, gaelico scozzese, inglese, irlandese, **italiano**, norvegese, olandese, portoghese, romancio, spagnolo, svedese e tedesco.
- Copre anche:
  - albanese, indonesiano, afrikaans e swahili.



| | _0 | _1 | _2 | _3 | _4 | _5 | _6 | _7 | _8 | _9 | _A | _B | _C | _D | _E | _F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0_ | | ┌ | ┐ | └ | ┘ | │ | – | • | ◘ | | | ♂ | □ | | ♫ | ☼ |
| 1_ | ┼ | ◄ | ↕ | ‼ | ¶ | ┴ | ┬ | ┤ | ↑ | ├ | → | ← | | | | |
| 2_ | | ! | " | # | $ | % | & | ' | ( | ) | * | + | , | - | . | / |
| 3_ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; | < | = | > | ? |
| 4_ | @ | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
| 5_ | P | Q | R | S | T | U | V | W | X | Y | Z | [ | \ | ] | ^ | _ |
| 6_ | ` | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o |
| 7_ | p | q | r | s | t | u | v | w | x | y | z | { | \| | } | ~ | |
| 8_ | € | | ‚ | ƒ | „ | … | † | ‡ | ˆ | ‰ | Š | ‹ | Œ | | Ž | |
| 9_ | | ‘ | ’ | “ | ” | • | – | — | ˜ | ™ | š | › | œ | | ž | Ÿ |
| A_ | | ¡ | ¢ | £ | ¤ | ¥ | ¦ | § | ¨ | © | ª | « | ¬ | - | ® | ¯ |
| B_ | ° | ± | ² | ³ | ´ | µ | ¶ | · | ¸ | ¹ | º | » | ¼ | ½ | ¾ | ¿ |
| C_ | À | Á | Â | Ã | Ä | Å | Æ | Ç | È | É | Ê | Ë | Ì | Í | Î | Ï |
| D_ | Ð | Ñ | Ò | Ó | Ô | Õ | Ö | × | Ø | Ù | Ú | Û | Ü | Ý | Þ | ß |
| E_ | à | á | â | ã | ä | å | æ | ç | è | é | ê | ë | ì | í | î | ï |
| F_ | ð | ñ | ò | ó | ô | õ | ö | ÷ | ø | ù | ú | û | ü | ý | þ | ÿ |

# ISO 8859

- Lo standard **ISO 8859** è complessivamente composto da 16 parti, ciascuna delle quali è progettata per rappresentare lingue simili, in modo che i comuni caratteri utilizzati siano inseriti nella stessa raccolta.
- Quando un simbolo è ripetuto in più parti, generalmente mantiene la stessa posizione, in modo da limitare i problemi di conversione.
- Oltre all'ISO Latin 1, è molto usato anche l'**ISO Latin 15** che lo ha sostituito. Nella revisione:
  - Sono stati eliminati alcuni simboli scarsamente utilizzati.
  - Questi simboli sono stati sostituiti con il simbolo dell'euro **€** e con le lettere Š, š, Ž, ž, Œ, œ, e Ÿ, che completano la copertura di francese, finlandese ed estone.

# Unicode e ISO/IEC 10646

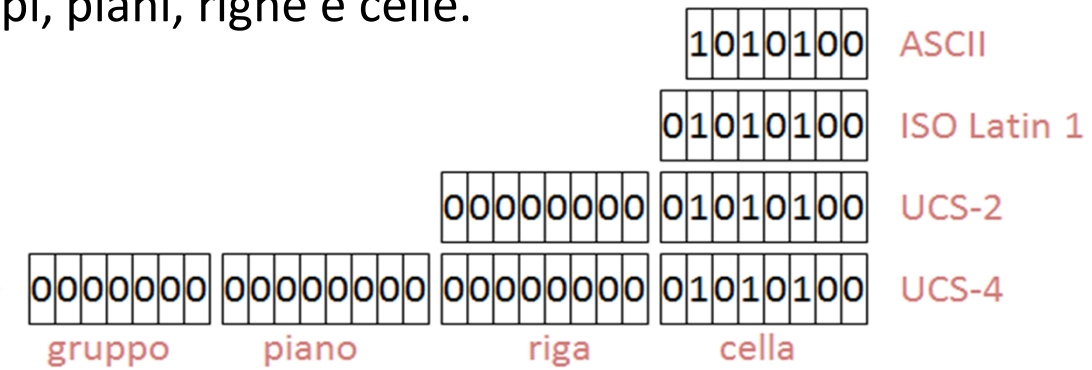
- ISO 8859 non risolve tutti i problemi legati alle lingue con alfabeti non latini (arabo, cinese, giapponese, tailandese, ecc).
- Per affrontare in modo definitivo le questioni di internazionalizzazione si mettono al lavoro due gruppi (uno di origine commerciale, uno di origine istituzionale), che producono due standard **Unicode** e **ISO/IEC 10646.**


- I due standard sono mantenuti sincronizzati dal 1991 ma in teoria questo sodalizio potrebbe rompersi e i due standard potrebbero procedere autonomamente.

# ISO/IEC 10646

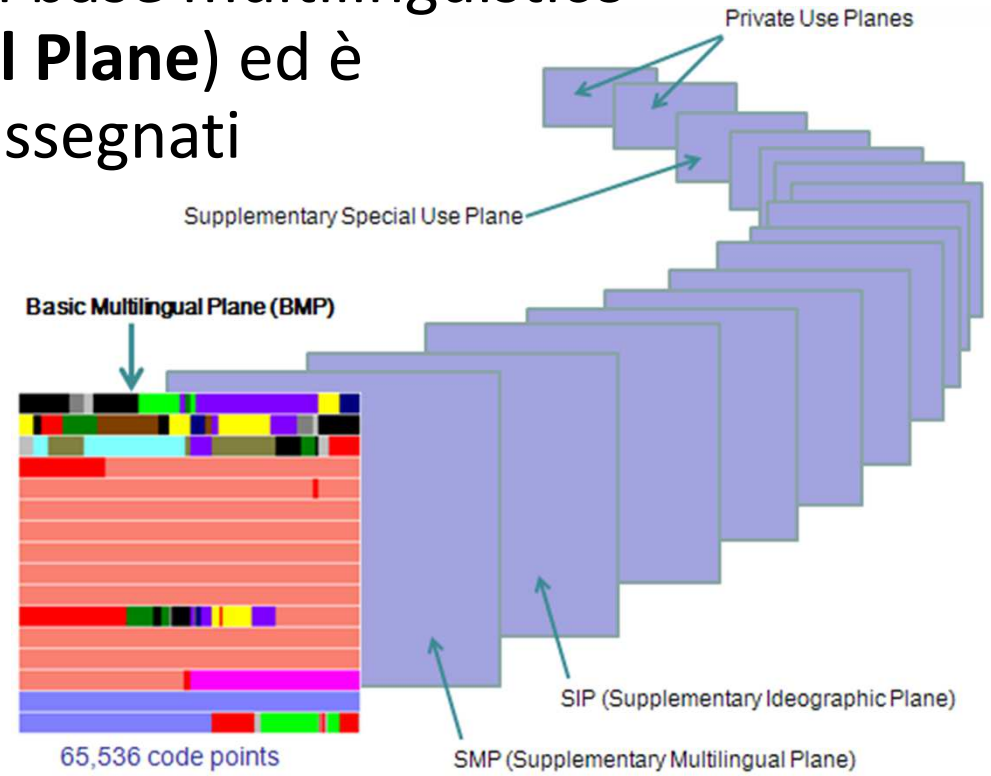
- ISO/IEC 10646 definisce:
  - 128 gruppi (groups) di
  - 256 piani (planes) di
  - 256 righe (rows) odif
  - 256 celle (cells)

  che potenzialmente identificano 2.147.483.648 caratteri (in realtà può codificare 679,477,248 caratteri)
- ISO 10646 è composto di due schemi di codifica.
  - UCS-2 è uno schema a due byte, che è un'estensione di ISO Latin 1.
  - UCS-4 è uno schema a 31 bit in 4 byte, estensione di UCS-2. E' diviso in gruppi, piani, righe e celle.



# ISO/IEC 10646: piani

- Dei 17 piani, ciascuno in grado di codificare 65.536 caratteri, sono assegnati solo i primi 3 e gli ultimi tre.
- Il piano 0 è detto piano di base multilinguistico (**BMP** - **Basic Multilingual Plane**) ed è il piano in cui sono stati assegnati la maggior parte dei caratteri.
- BMP contiene caratteri per quasi tutti i moderni linguaggi e un grande numero di caratteri speciali



# UCS e UTF

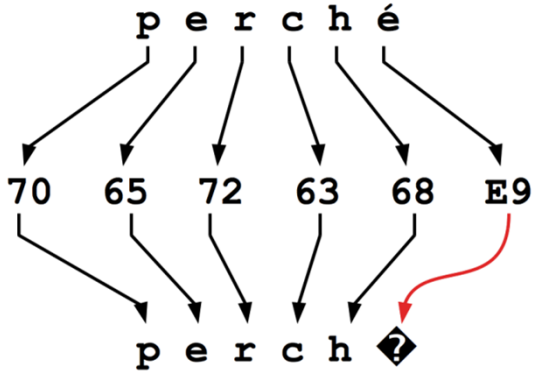
- Unicode e ISO/IEC 10646 utilizzano 4 byte per la codifica di un solo carattere:
  - Risolvono i problemi di codifica delle lingue non europee MA
  - Consumano molta memoria
- In realtà la maggior parte degli alfabeti sta nel BMP, e la maggior parte dei documenti sono scritti in ASCII.
- Quindi, al posto di UCS si usa **UTF** (**UCS Transformation Format**), che consente di usare tutti i caratteri definiti in UCS ma utilizzando una codifica a lunghezza variabile.

# UTF-16 e UTF-8

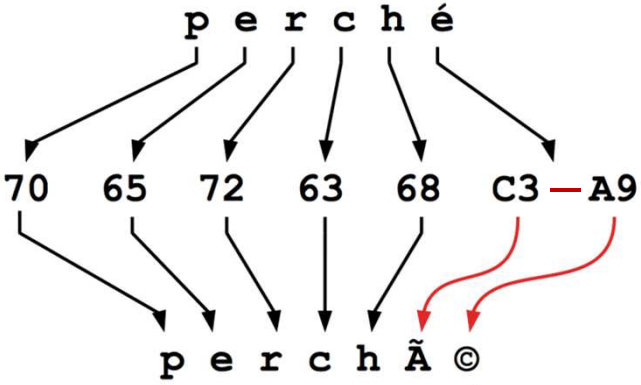
- UTF-16: considera tutti i caratteri di UCS-2 (o in 16 bit).
- UTF-8 considera di accedere a tutti i caratteri di UCS-4, ma utilizza un numero compreso tra 1 e 4 byte per farlo.
  - I codici compresi tra 0 - 127 (ASCII a 7 bit), e richiedono un byte (sempre 0 al primo bit).
  - I codici derivati dall'alfabeto latino e tutti gli script non-ideografici richiedono 2 byte.
  - I codici ideografici (orientali) richiedono 3 byte
  - I codici dei piani alti richiedono 4 byte.

# UTF-8 e Latin 1

- Le codifiche UTF-8 e Latin 1 sono compatibili ma non identiche
- Ci possono essere problemi:
  - Aprendo un Latin 1 come UTF8:




  - Aprendo un UTF8 come Latin 1:



# Riferimenti

- Libro: capitolo 9 (Text and Typography)
- Risorse on line sulla piattaforma
- Standard di riferimento