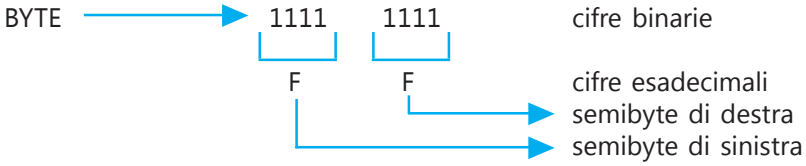
Codice ASCII e codice Unicode

Informazioni alfanumeriche

Con un bit si hanno due combinazioni possibili: 0, 1. Con un byte abbiamo 256 combinazioni possibili: da 00000000 a 11111111.
Infatti 11111111 in base 2 = 128 + 64 + 32 + 16 + 8 + 4 + 2 + 1 = 255.
Il massimo valore che può esprimere un byte è 255 decimale.
Il massimo numero di combinazioni che può contenere un byte è 256 (compresa la combinazione 00000000).
Un byte nel quale tutti i bit sono uguali a 1 (11111111) è equivalente a FF esadecimale.
Per rappresentare un simbolo del sistema esadecimale occorrono 4 bit (4 cifre binarie).



La rappresentazione di una cifra esadecimale richiede un *semibyte*.
Tutte le informazioni non numeriche (alfabetiche o alfanumeriche) sono esprimibili mediante una combinazione di lettere, cifre o caratteri speciali: affinché un elaboratore riesca a riconoscere e a trattare tali informazioni deve essere stabilita una corrispondenza che ad ogni carattere utilizzato per rappresentare le informazioni associ una particolare configurazione degli 8 bit di un byte.

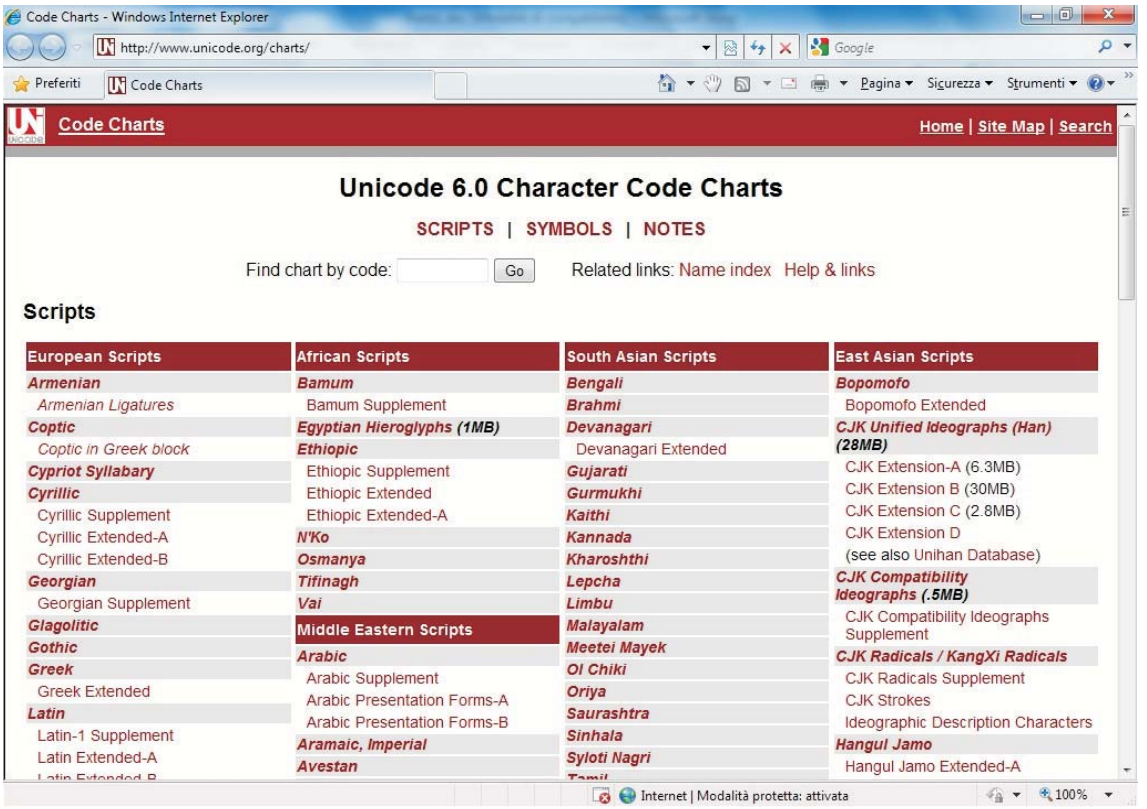
> La rappresentazione dei dati all'interno di un elaboratore è quindi realizzata attraverso l'associazione di una combinazione di bit ad un determinato simbolo (lettera, cifra o carattere speciale): questa associazione è chiamata **codifica**.

La codifica di base per i caratteri di un testo si chiama **ASCII** (*American Standard Code for Information Interchange*, in italiano *codice americano standard per lo scambio di informazioni*) che utilizza 7 bit per codificare un singolo carattere; per esempio a "0110000" corrisponde il carattere "0" (cifra zero) o a "1110001" corrisponde la lettera "q" (minuscola).
I 7 bit del codice consentono di rappresentare 128 (2^7) caratteri diversi tra loro e quindi non sono sufficienti per codificare tutte le variazioni e tutti i simboli utilizzati nelle varie lingue. Questo sistema è stato esteso a 8 bit, raddoppiando il numero di caratteri disponibili (256 = 2^8), con la definizione del codice chiamato **ASCII esteso**.

Codice ASCII: primi 127 caratteri	
Da 1 a 31 e 127	Caratteri non stampabili, ma che hanno un significato particolare, per esempio *carattere vuoto*, *fine del testo*, *segnale acustico*, *Cancel*, *Esc*.
Da 32 (0010 0000) a 47	Il carattere spazio (32) e i segni di punteggiatura quali: punto esclamativo, punto, virgola.
Da 48 (0011 0000) a 57	Cifre da 0 a 9.
Da 58 (0010 0000) a 64	Altri segni di punteggiatura quali: due punti, maggiore, uguale, minore.
Da 65 (0100 0001) a 90	Lettere maiuscole da "A" a "Z".
Da 91 (0101 1011) a 96	Altri segni di punteggiatura: per esempio, apostrofo, parentesi quadre.
Da 97 (0110 0001) a 122	Lettere minuscole da "a" a "z".
Da 123 (0111 1011) a 126	Altri segni di punteggiatura: per esempio, parentesi graffe, tilde.

Nel 1991 è stata sviluppata una nuova codifica, chiamata **Unicode**, che si è diffusa rapidamente ed è in continua evoluzione: è diventata lo standard di fatto per la rappresentazione delle informazioni nei documenti elettronici, in particolare nelle pagine Web, utilizzando i simboli delle numerose lingue esistenti nel mondo.

Le tabelle dei codici *Unicode* sono disponibili sul sito *http://www.unicode.org/charts*.



I primi caratteri di *Unicode* sono esattamente gli stessi della codifica ASCII, in modo da mantenere la compatibilità con il sistema preesistente. All'inizio la codifica utilizzava 2 byte (16 bit, con la possibilità di codificare 65.536 caratteri), ma poi è stata estesa a 32 bit, permettendo la rappresentazione di più di un milione di caratteri differenti.

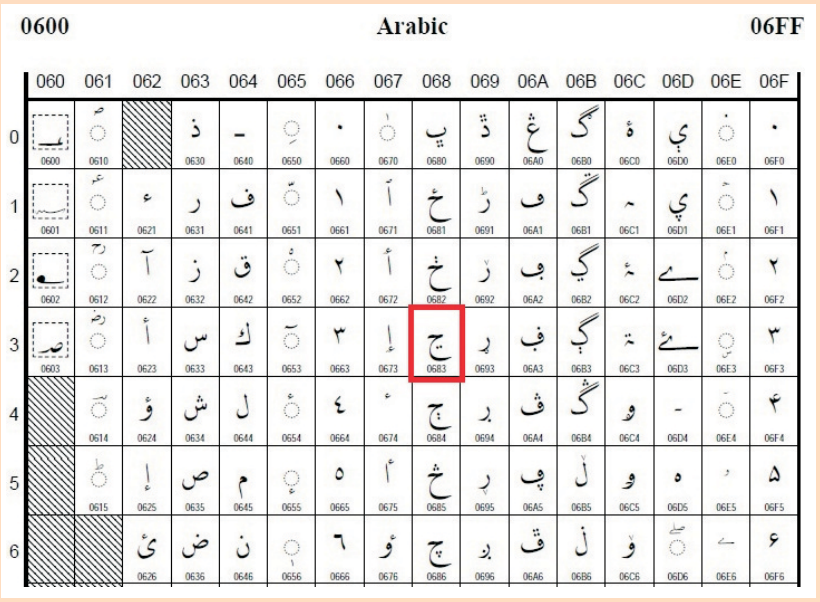
L'obiettivo generale di *Unicode* è di creare una codifica che comprenda tutti i caratteri, con tutte le variazioni possibili, di tutte le lingue esistenti, oltre ai simboli utilizzati in matematica e nelle scienze.

Per semplificare le operazioni sono state poi create versioni ridotte del codice che permettono di scrivere i caratteri di uso più frequente in modo più breve: **UTF-8** (a 8 bit), **UTF-16** (a 16 bit) e **UTF-32** (a 32 bit).

La stringa di testo "*Ciao, mondo!*" contenente 12 caratteri (occorre tenere conto di tutti i simboli, compresi la virgola, lo spazio e il punto esclamativo) occuperebbe 84 bit (12 x 7) se codificata in *ASCII standard*, mentre ne occuperebbe 384 (12 x 32) in *UTF-32*.

ESEMPIO

La figura seguente mostra la prima parte del documento *Unicode* per la codifica dei caratteri delle lingue arabe (codice *Arabic*): i caratteri corrispondono ai numeri compresi nell'intervallo da *0600* a *06FF* in esadecimale.



Il carattere evidenziato con un riquadro rosso corrisponde al codice 0683 in esadecimale (colonna 068, riga 3 della tabella).

ESEMPIO

Inserire in un documento Word un carattere Unicode.
Volendo inserire il carattere della figura in un documento Word, occorre convertire il valore in esadecimale della tabella Unicode in numero decimale (si può usare la *Calcolatrice* di Windows in *Accessori*):

$$0683_{16} = 1667_{10}$$

Poi, in Word, occorre tenere premuto il tasto *Alt* e premere in successione i tasti *1667* sul tastierino numerico (a destra nella tastiera).
In alternativa, si possono scrivere direttamente nel testo le cifre esadecimali del codice Unicode e premere subito dopo la combinazione di tasti **Alt + X**.
Per esempio, scrivendo le cifre 20AC seguite da Alt+X si ottiene il simbolo dell'euro €.