Rappresentazione di testi

Il codice ASCII

- Per rappresentare gli elementi di un testo
  - caratteri minuscoli e maiuscoli (52)
  - numeri (10)
  - segni di punteggiatura e altri simboli grafici
  - caratteri di controllo (spazio, a capo, linea seguente, tabulatore)

  si può creare una corrispondenza CONVENZIONALE con i primi N numeri binari

- Negli anni si è affermata la codifica ASCII, mostrata in tabella. - Ancora largamente usata nei piccoli sistemi in cui sono elaborati testi (p. es. DISPLAY alfanumerici).

  American Standard Code for Information Interchange
  (1960 Bell Labs; pubblicato nel 1963)

- Osservazioni
  - numeri e lettere sono "in ordine"
    (facilita le operazioni di ordinamento o "sort")
  - I numeri 0..9 vanno da $\emptyset x 3\emptyset$ a $0x39$
    Per convertire da CODICE a VALORE NUMERICO basta considerare i 4 bit meno significativi
  - Le lettere MAIUSCOLE vanno da $\emptyset x41$ ('A') a $\emptyset x5A$ ('Z')
    e precedono tutte le minuscole che vanno da
    $\emptyset x61$ ('a') a $\emptyset x8A$ ('z')

    Quindi per passare da maiuscolo a minuscolo si somma $\emptyset x2\emptyset$ (cioè 32) e si sottrae per il viceversa.
  - I caratteri "di controllo" vanno da $\emptyset x\emptyset\emptyset$ a $\emptyset x2\emptyset$ e inoltre c'è $0xFF$ e sono indicati da 2 o 3 lettere maiuscole (abbreviazioni o acronimi) che ne indicano il significato.
  - Non ci sono elementi di formattazione

#7.2

USASCII code chart

Bits b_7 b_6 b_5 →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b_4 b_3 b_2 b_1 ↓ Row / Column →	0	1	2	3	4	5	6	7
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	A 10	LF	SUB	*	:	J	Z	j	z
1 0 1 1	B 11	VT	ESC	+	;	K	[	k	{
1 1 0 0	C 12	FF	FS	,	<	L	\	l	|
1 1 0 1	D 13	CR	GS	-	=	M	]	m	}
1 1 1 0	E 14	SO	RS	.	>	N	^	n	~
1 1 1 1	F 15	SI	US	/	?	O	_	o	DEL

BINARIO

HEX

BINARIO

HEX

Binary	Oct	Dec	Hex	Abbr.	Name
000 0000	0	0	00	NUL	Null
000 0001	1	1	01	SOH	Start of Heading
000 0010	2	2	02	STX	Start of Text
000 0011	3	3	03	ETX	End of Text
000 0100	4	4	04	EOT	End of Transmission
000 0101	5	5	05	ENQ	Enquiry
000 0110	6	6	06	ACK	Acknowledgement
000 0111	7	7	07	BEL	Bell
000 1000	10	8	08	BS	Backspace
000 1001	11	9	09	HT	Horizontal Tab
000 1010	12	10	0A	LF	Line Feed
000 1011	13	11	0B	VT	Vertical Tab
000 1100	14	12	0C	FF	Form Feed
000 1101	15	13	0D	CR	Carriage Return[h]
000 1110	16	14	0E	SO	Shift Out
000 1111	17	15	0F	SI	Shift In
001 0000	20	16	10	DLE	Data Link Escape
001 0001	21	17	11	DC1	Device Control 1 (often XON)
001 0010	22	18	12	DC2	Device Control 2
001 0011	23	19	13	DC3	Device Control 3 (often XOFF)
001 0100	24	20	14	DC4	Device Control 4
001 0101	25	21	15	NAK	Negative Acknowledgement
001 0110	26	22	16	SYN	Synchronous Idle
001 0111	27	23	17	ETB	End of Transmission Block
001 1000	30	24	18	CAN	Cancel
001 1001	31	25	19	EM	End of Medium
001 1010	32	26	1A	SUB	Substitute
001 1011	33	27	1B	ESC	Escape
001 1100	34	28	1C	FS	File Separator
001 1101	35	29	1D	GS	Group Separator
001 1110	36	30	1E	RS	Record Separator
001 1111	37	31	1F	US	Unit Separator
111 1111	177	127	7F	DEL	Delete

UNICODE

U+HHH valore esadecimale

- L'evoluzione dei sistemi per la gestione di testi ha richiesto anche un aggiornamento del sistema di codifica dei caratteri. Uno dei più interessanti è unicode (www.unicode.org)
  - Allineato a ISO/IEC 10646 (standard internazionale)
  - Universale
  - inglobe ASCII (i primi 128 codici)
  - prevede diverse forme di codifica (UTF-8, UTF-16, UTF-32); include potenzialmente oltre 10^6 codici

> Comprende un data-base ordinato di caratteri (elementi di testo) simboli, segni grafici, modificatori ...

> include il modo di tradurli in stream di quantità binarie (encoding)

> specifica l'identità del carattere, NON IL GLIFO (la forma grafica - font-specifica)

NE' ALTRI ELEMENTI GRAFICI (colore, dim., enfasi ...)

> Prevede la costruzione di caratteri composti (ü; é ...) o con un codice specifico ü = U+00FC

oppure con i "non-spacing characters"

ü = U+0075; U+0308
(u) (es dieresi)

> Permette di specificare la DIREZIONE (Sx → Dx ; Dx → Sx) e seguire l'ordine logico

> Univocità del "CODE POINT" (evitare duplicazione di codici per lo stesso carattere) associato all'identificatore del carattere

> Range di codici definiscono un "CODE SPACE" lo standard cerca di rispettare l'ordine tradizionale dei caratteri di una scritta

> STABILE: non vengono modificati codici già assegnati, ma solo aggiunti

UTF-8

- Una delle possibili codifiche dei caratteri unicode (Ken Thompson, Rob Pike 1993)
- Codifica a lunghezza variabile
  Schema di codifica (dal 2003, limitati ai set UNICODE)

BYTE #	bit UTILI	Range U+ hex		1°	2°	3°	4°
1	7	0	7F	0xxxxxxx	→ codici ASCII		
2	11	80	7FF	110xxxxx	10xxxxxx		
3	16	800	FFFF	1110xxxx	10xxxxxx	10xxxxxx	
4	21	10000	10 FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

Le x di peso più grande di quelle richieste dal codice sono riempite con 0. Si deve usare la codifica con meno byte.

Vantaggi

- Retrocompatibile con ASCII
- Permette l'individuazione euristica con buon successo (esaminando i primi bit di ogni byte) della codifica
- Chiara distinzione tra singolo byte e multi byte (INIZIANO con 1)
- Chiara indicazione del numero di byte di una sequenza (gli 1 iniziali del primo byte)