Linguistica Computazionale

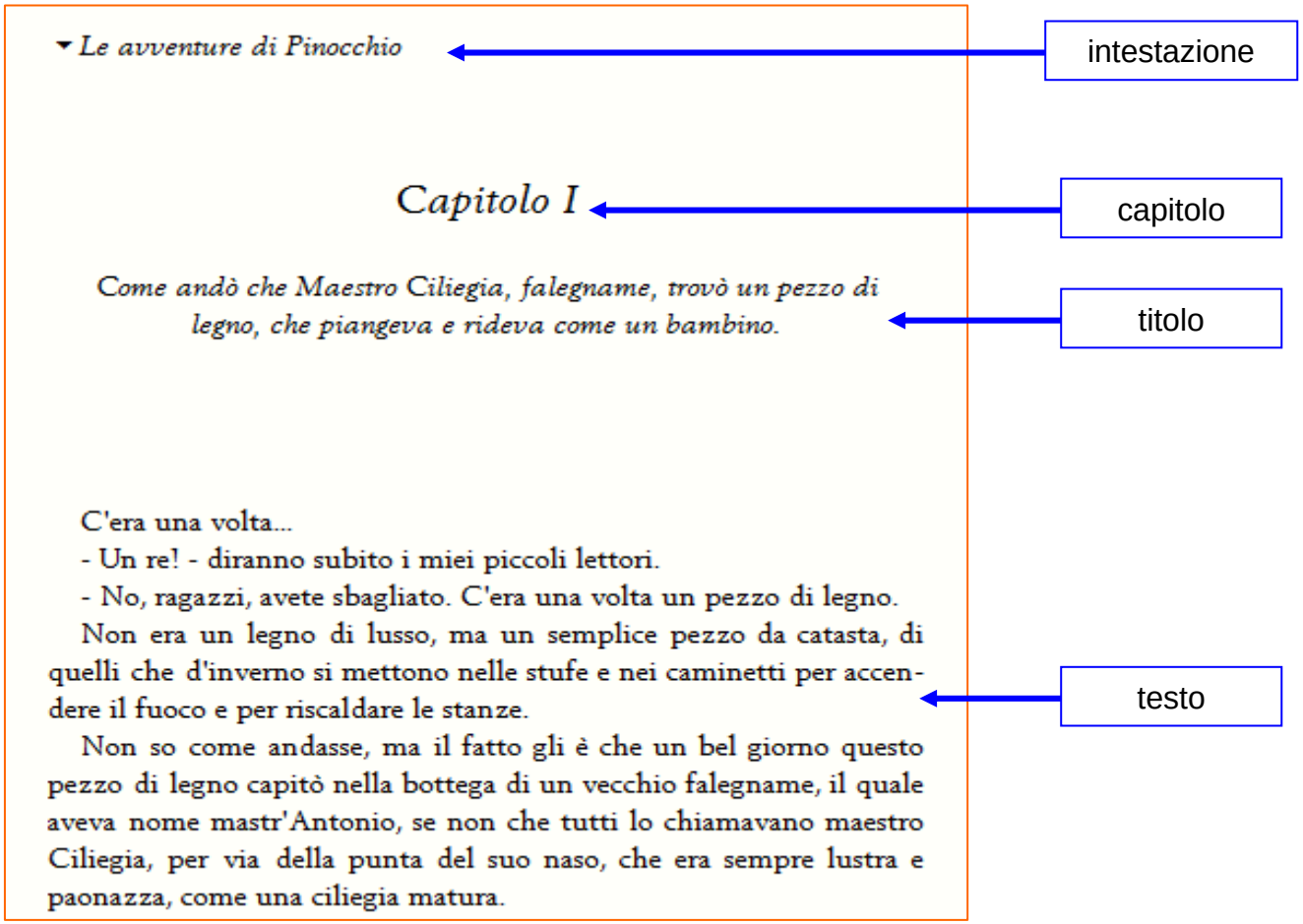
La codifica digitale del testo

Salvatore Sorce
Dipartimento di Ingegneria
Chimica, Gestionale, Informatica e Meccanica

Lucidi Adattati da Alessandro Lenci
Dipartimento di Linguistica “T. Bolelli”

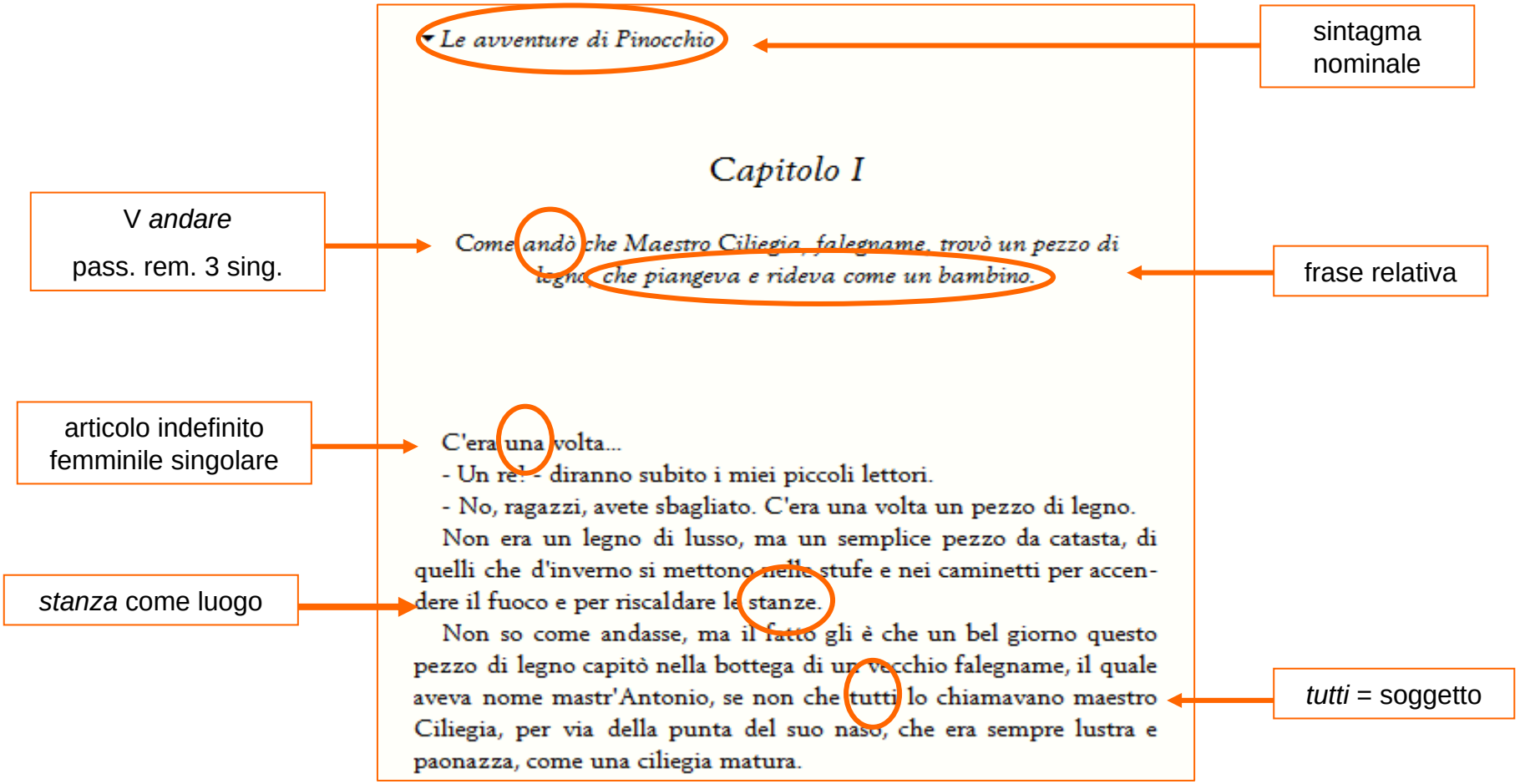
La codifica digitale del testo

Il testo e la sua organizzazione



La codifica digitale del testo

Il testo e la sua struttura linguistica



La codifica digitale del testo

Rappresentazione del testo su supporto digitale in un formato “leggibile” da un computer

Machine Readable Form (MRF)

- I computer memorizzano ed elaborano dati sotto forma di sequenze di due soli simboli 0 e 1 (cifre binarie)
- Ogni tipo di informazione deve essere codificata in cifre binarie
  - codificare informazione = associare a ciascuna unità di informazione un codice (sequenza di cifre binarie) che la identifica in maniera univoca
- I testi per essere elaborati o trasmessi da un programma devono avere una rappresentazione (codifica) binaria

La codifica digitale del testo

Il testo come sequenza di caratteri

Ciascun carattere alfanumerico, di punteggiatura o di controllo che compone il testo deve essere rappresentato nei termini di un codice binario

```
Le avventure di Pinocchio
Capitolo I
Come andò che Maestro Ciliegia, falegname, trovò un pezzo di legno, che
piangeva e rideva come un bambino.
C'era una volta...
- Un re! - diranno subito i miei piccoli lettori.
- No, ragazzi, avete sbagliato. C'era una volta un pezzo di legno.
Non era un legno di lusso, ma un semplice pezzo da catasta, di quelli che
d'inverno si mettono nelle stufe e nei caminetti per accendere il fuoco e per
riscaldare le stanze.
Non so come andasse, ma il fatto gli è che un bel giorno questo pezzo di
legno capitò nella bottega di un vecchio falegname, il quale aveva nome
mastr'Antonio, se non che tutti lo chiamavano maestro Ciliegia, per via della
punta del suo naso, che era sempre lustra e paonazza, come una ciliegia
matura.
```

La codifica digitale del testo

Il testo come sequenza di caratteri

- "Surrogato" parziale del testo originario
  - completa equivalenza solo dal punto di vista dei caratteri che lo compongono
  - perdita di informazione
    - l'informazione implicitamente veicolata dalla formattazione del testo relativa a:
      - le coordinate meta-testuali
        - » il nome dell'autore, il titolo, ecc.
      - la struttura e organizzazione testuale
        - » la suddivisione logica in sezioni, capitoli, paragrafi, ecc.
  - nessun guadagno di informazione
    - l'informazione sulla struttura linguistica rimane implicita e nascosta (come nel testo originale)

La codifica digitale del testo

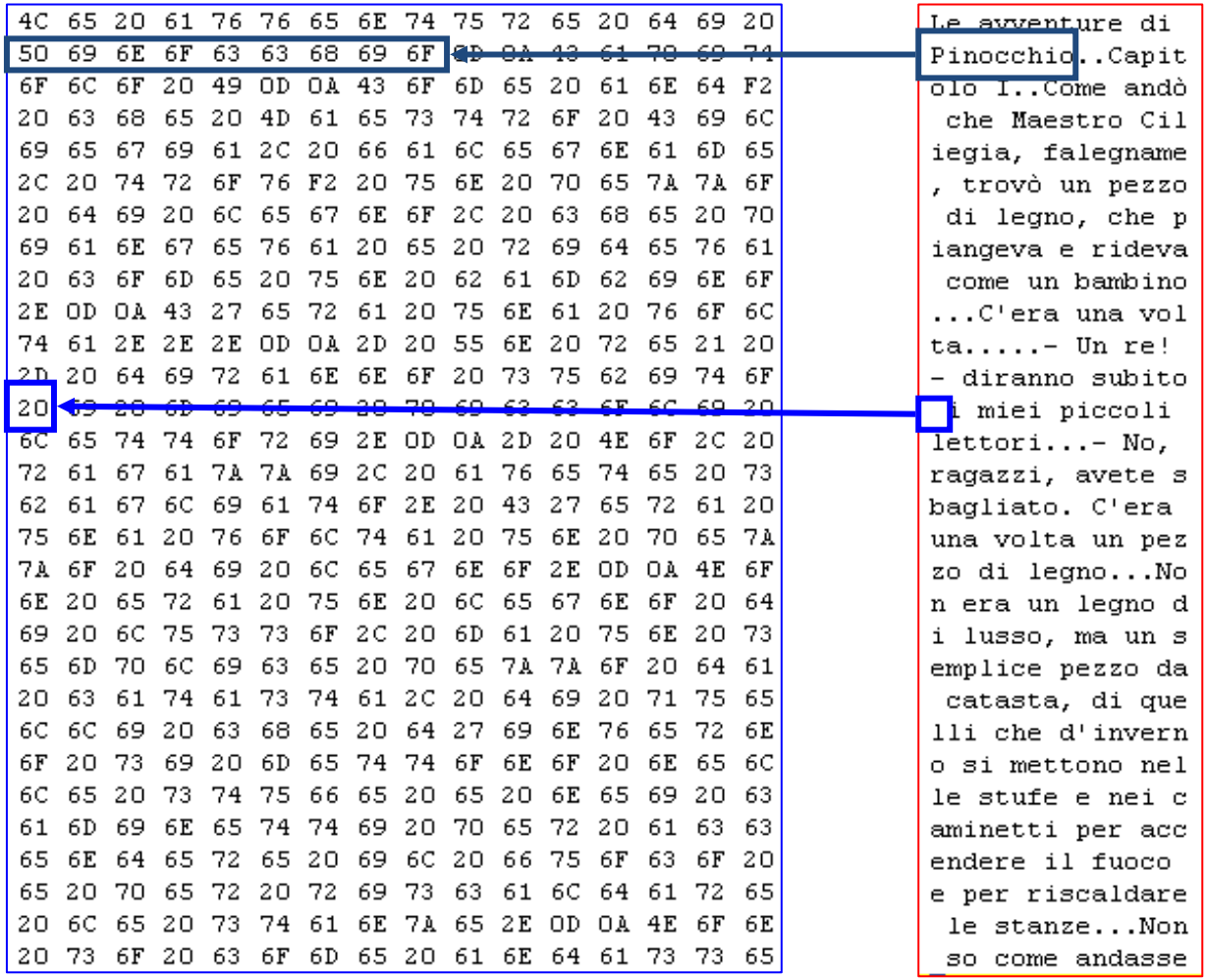
- Due livelli di codifica del testo digitale
  - codifica di basso livello (codifica di livello 0)
    - riguarda la rappresentazione binaria della sequenza ordinata dei caratteri
  - codifica di alto livello
    - arricchisce il testo codificato al livello zero con informazione relativa a dimensioni strutturali
      - organizzazione del testo in strutture macrotestuali
      - articolazione del testo in strutture linguistiche

La codifica di alto livello permette di rendere esplicita qualsiasi interpretazione, anche di tipo linguistico, si voglia associare al testo

La codifica di livello 0

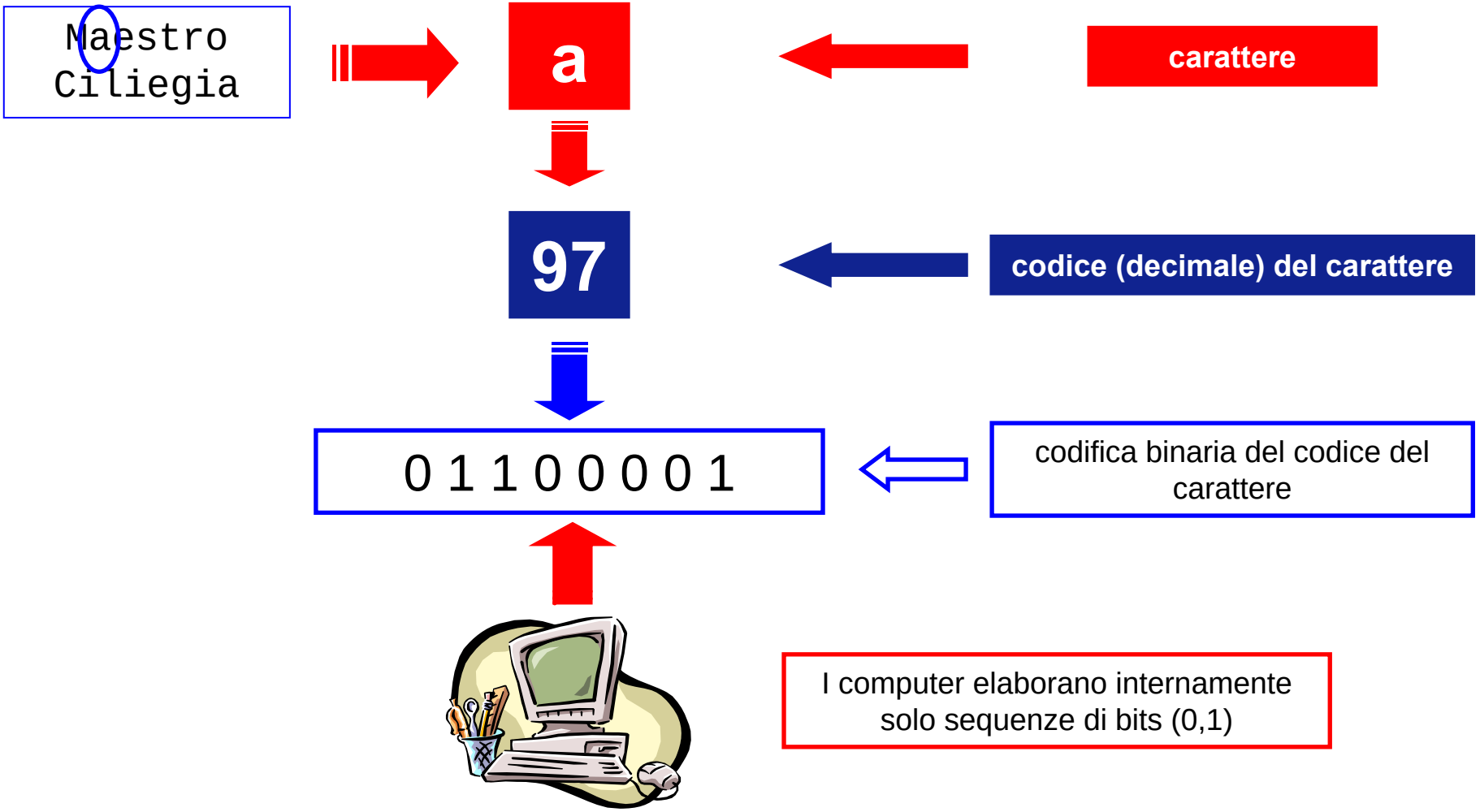
Il testo come sequenza di caratteri

dietro le quinte…

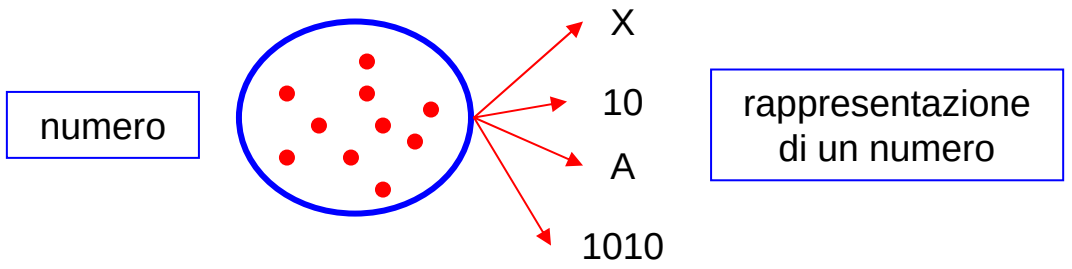


La codifica di livello 0

caratteri e numeri



Numeri e numeri



- Sistema binario
  - vengono usate due cifre (0 e 1) per rappresentare un numero
  - problema: i numeri binari sono estremamente lunghi e difficili da ricordare
- Sistema esadecimale
  - ogni numero è rappresentato con 16 cifre (0-9, A-F)
  - i numeri sono più corti di quelli binari
  - estrema facilità di conversione tra binario ed esadecimale
    - in una sequenza binaria, ogni stringa di 4 bits corrisponde ad una cifra esadecimale
    - 0110 1111 0110 numero binario
    - (6) (15) (6)
    - 6 F 6 numero esadecimale

Come sono rappresentati i caratteri nel computer?

- Repertorio di caratteri
  - un insieme di caratteri (es. “A”, “a”, “!”, “à”, “P”, ecc.)
  - i caratteri sono entità astratte, da non confondersi con il modo in cui sono realizzati graficamente (gliphs)
    - “*a*”, “**a**”, “a”, “a” sono tutti lo stesso carattere “a”
    - la stessa realizzazione grafica può corrispondere a caratteri diversi (es. “A” latino e “A” cirillico e “A” greco)
- Set di carattere (codice)
  - una tabella che definisce una corrispondenza biunivoca (1-a-1) tra un repertorio di caratteri e un insieme di numeri interi non negativi
    - a ogni carattere è assegnato un codice numerico (punto di codice o code position)
- Codifica di carattere
  - algoritmo che determina come i codici dei caratteri sono rappresentati in sequenze di bits (bytes)

Il codice ASCII

- Primo standard per l'assegnazione di codici a caratteri (dal 1963)
  - set di caratteri riconosciuto da tutti i computer
  - conosciuto come "ASCII Standard" o ISO-646
- Codifica
  - 7 bits
    - ciascun punto di codice è rappresentato con il numero binario corrispondente di 7 bits
    - in realtà 1 byte = 8 bits di cui un bit non è usato per la codifica (bit di parità)
  - 7 bits = 2^7 punti di codice = 128 caratteri rappresentati
- Sufficiente per rappresentare l'inglese
  - mancano i caratteri accentati, ecc. per rappresentare altri alfabeti occidentali

ASCII Standard

decimale ed esadecimale

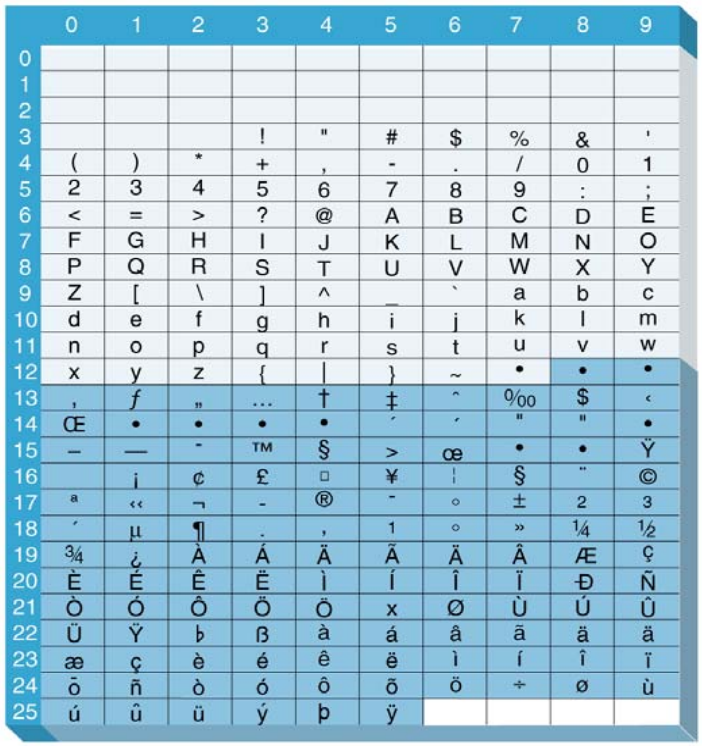
Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.asciitable.com

Il set di caratteri ISO-Latin-1

- ISO-Latin-1 (ISO-8859-1 o ASCII esteso)
  - unica estensione standard di ASCII
  - 1 byte = 8 bits = 2^8 punti di codice = 256 caratteri rappresentati
  - sufficiente per lingue europee occidentali (italiano, francese, ecc.)

ASCII Standard



Caratteri di controllo

0-32

128-159

La famiglia di caratteri ISO-8859

- 14 set di caratteri standardizzati da ISO (International Standard Organization)
- Codifica
  - 1 byte = 256 caratteri rappresentati da ciascun set
- Soprainsiemi dei caratteri ASCII Standard
  - punti di codice 0 - 127 (parte comune) ASCII
  - punti di codice 128 - 159 codici di controllo (non corrispondono a caratteri grafici)
  - punti di codice 160 - 255 (parte variabile) caratteri aggiuntivi per greco, cirillico, lingue slave, arabo, ecc.
- I set di ISO-8859 sono tutti reciprocamente incompatibili
  - Punto di codice 232
    - ISO-8859-1 (Latin-1) = "è"
    - ISO-8859-6 (Cyrillic) = "ш"
- ISO-8859 non copre lingue come giapponese, cinese, ecc.

La famiglia di caratteri ISO-8859

ISO-Latin-1

The parts of ISO 8859

standard	name of alphabet	characterization
ISO 8859-1	Latin alphabet No. 1	"Western", "West European"
ISO 8859-2	Latin alphabet No. 2	"Central European", "East European"
ISO 8859-3	Latin alphabet No. 3	"South European"; "Maltese & Esperanto"
ISO 8859-4	Latin alphabet No. 4	"North European"
ISO 8859-5	Latin/Cyrillic alphabet	(for Slavic languages)
ISO 8859-6	Latin/Arabic alphabet	(for the Arabic language)
ISO 8859-7	Latin/Greek alphabet	(for modern Greek)
ISO 8859-8	Latin/Hebrew alphabet	(for Hebrew and Yiddish)
ISO 8859-9	Latin alphabet No. 5	"Turkish"
ISO 8859-10	Latin alphabet No. 6	"Nordic" (Sámi, Inuit, Icelandic)
ISO 8859-11	Latin/Thai alphabet	(for the Thai language)
(Part 12 has not been defined.)		
ISO 8859-13	Latin alphabet No. 7	Baltic Rim
ISO 8859-14	Latin alphabet No. 8	Celtic
ISO 8859-15	Latin alphabet No. 9	"euro"
ISO 8859-16	Latin alphabet No. 10	for a collection of languages (see below)

The Universal Character Set

UNICODE (ISO-10646)

- Standard internazionale che permette di rappresentare qualsiasi tipo di carattere appartenente ai sistemi grafici esistenti
  - lingue europee, asiatiche, arabo, ebraico, cirillico, ugaritico, ecc.
  - basato su principi di composizione dinamica dei caratteri utile per caratteri complessi, e.g., cinesi latini con segni diacritici, etc.
    - ç = c + ¸
- Assegna un numero di codice univoco ad ogni carattere
  - "è" = 232
  - "Ш" = 1096
- Risolve i problemi di incompatibilità dei sistemi ISO-8859
  - estende l'insieme dei caratteri supportati
  - permette la realizzazione di documenti multilingue
- Unicode è un soprainsieme di ASCII
- http://www.unicode.org

The Universal Character Set

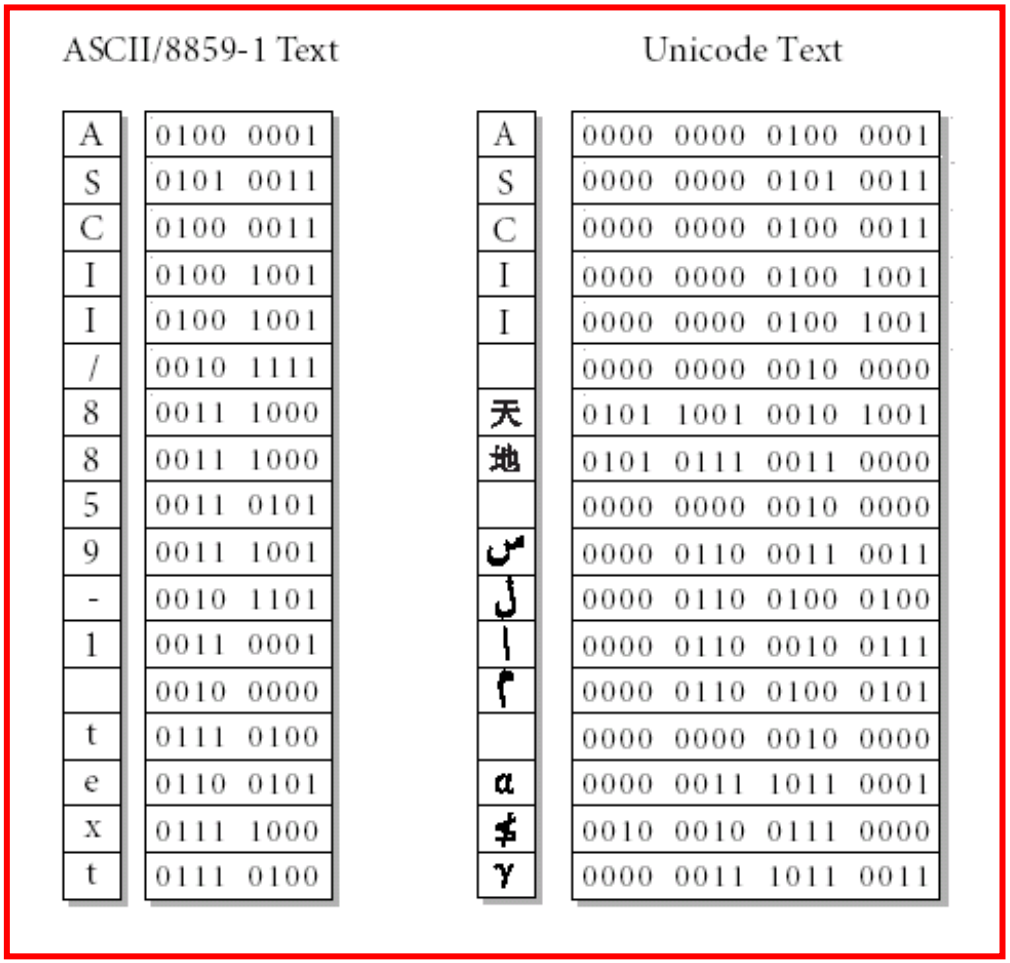
UNICODE (ISO-10646)

- Molteplici tipi di codifica:
  - UCS-2, UCS-4, UTF-8, UTF-16, ecc.
  - Codifica comune UTF-8
    - codifica di Unicode a lunghezza variabile che usa da 1 a 4 bytes per ogni carattere
    - UTF-8 usa 1 byte per la codifica dei caratteri corrispondenti al set ASCII (cioè la compatibilità non si estende per i caratteri da 128 a 255)
    - totale compatibilità con la codifica ASCII (…ma non con ISO-latin –1!!!)

- Arabic
- Armenian
- Bengali
- Bopomofo
- Buhid
- Canadian Syllabics
- Cherokee
- Cyrillic
- Deseret
- Devanagari
- Ethiopic
- Georgian
- Gothic
- Greek
- Gujarati
- Gurmukhi
- Han
- Hangul
- Hanunóo
- Hebrew
- Hiragana
- Kannada
- Katakana
- Khmer
- Latin
- Lao
- Malayalam
- Mongolian
- Myanmar
- Ogham
- Old Italic (Etruscan)
- Oriya
- Runic
- Sinhala
- Syriac
- Tagalog
- Tagbanwa
- Tamil
- Telugu
- Thaana
- Thai
- Tibetan
- Yi

The Universal Character Set

UNICODE (ISO-10646)



ASCII/8859-1 Text

A	0100 0001
S	0101 0011
C	0100 0011
I	0100 1001
I	0100 1001
/	0010 1111
8	0011 1000
8	0011 1000
5	0011 0101
9	0011 1001
-	0010 1101
1	0011 0001
	0010 0000
t	0111 0100
e	0110 0101
x	0111 1000
t	0111 0100

Unicode Text

A	0000 0000 0100 0001
S	0000 0000 0101 0011
C	0000 0000 0100 0011
I	0000 0000 0100 1001
I	0000 0000 0100 1001
	0000 0000 0010 0000
天	0101 1001 0010 1001
地	0101 0111 0011 0000
	0000 0000 0010 0000
س	0000 0110 0011 0011
ل	0000 0110 0100 0100
ا	0000 0110 0010 0111
م	0000 0110 0100 0101
	0000 0000 0010 0000
α	0000 0011 1011 0001
≤	0010 0010 0111 0000
γ	0000 0011 1011 0011

Caratteri e computer