

Editoria Elettronica

Vincenzo Gervasi

Dipartimento di Informatica

Università di Pisa

email: gervasi@di.unipi.it

www: <http://www.di.unipi.it/~gervasi>

Logistica del corso

- Orario delle lezioni:
 - Prima parte (Marzo, Aprile)
 - **Martedì 16:15-17:45 aula L**
 - **Mercoledì 16:15-17:45 aula L**
 - Seconda parte (Maggio, Giugno)
 - **Lezioni teoriche come sopra**
 - **3-4 lezioni in un laboratorio, date da confermare**
 - Probabilmente il martedì in laboratorio H

Logistica del corso

- Ricevimento studenti:
 - **Mercoledì 17:45 studio 333 (Dip. Inf.)**
 - **altro giorno/orario?**
- Per casi particolari:
 - su appuntamento, spedire email a gervasi@di.unipi.it
 - durante le pause delle lezioni

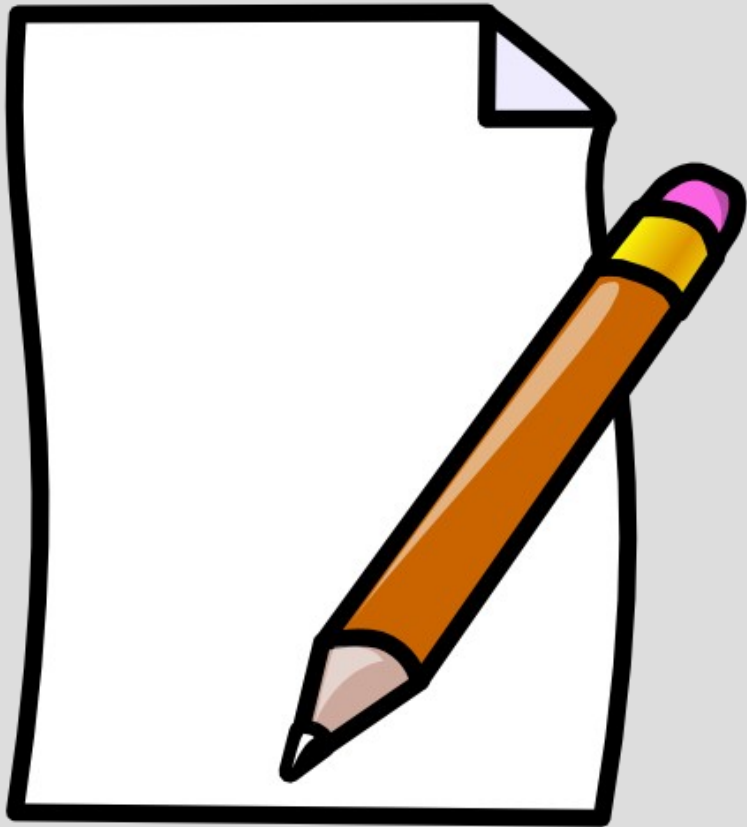
Logistica del corso

- Il corso (6 CFU) viene offerto a:
 - Studenti del Corso di Laurea magistrale in Informatica Umanistica (DM270)
 - Vecchi studenti del Corso di Laurea Specialistica in Informatica Umanistica (percorso “Editoria Elettronica”)
 - Vecchi studenti del Corso di Laurea in Informatica Umanistica (triennale)
 - Studenti di altri corsi di laurea (come esame libero)
 - **NON** è offerto a studenti della “nuova triennale” (DM270)

Logistica del corso

- Valutazione
 - prove in itinere (esercitazioni)
 - un elaborato finale (grafico o programma)
 - una prova orale (teorica)
- Materiale didattico
 - dispense del docente (lucidi, senza animazioni!)
 - altro materiale, anche online, verrà indicato di volta in volta durante il corso

Obiettivi del corso



- Formare **figure professionali** capaci di operare in **ambienti digitali** per la **gestione** e la **pubblicazione di informazioni strutturate e non strutturate** (testi e dati).

Obiettivi del corso

- Formare **figure professionali** capaci di operare in ambienti digitali per la gestione e la pubblicazione di informazioni strutturate e non strutturate (testi e dati).
- Il corso intende mettere i partecipanti in condizione di operare a livello professionale nel mondo dell'editoria
- Forniremo quindi molte informazioni di rilevanza pratica (e un bel po' di teoria)
- Parte del corso sarà dedicata all'acquisizione diretta di esperienze

Obiettivi del corso

- Formare figure professionali capaci di operare in **ambienti digitali** per la gestione e la pubblicazione di informazioni strutturate e non strutturate (testi e dati).
- La gran parte dell'attività editoriale tipica si svolge oggi con l'ausilio di **mezzi elettronici/digitali**
- La gestione di **informazioni e contenuti digitali** è parte essenziale della **società della conoscenza**
- L'intero **ciclo di vita dell'informazione** è ormai digitale!

Obiettivi del corso

- Formare figure professionali capaci di operare in ambienti digitali per la **gestione** e la **pubblicazione** di informazioni strutturate e non strutturate (testi e dati).
- **Non** ci occuperemo della **generazione** dell'informazione
 - giornalismo, scrittura creativa, scrittura tecnica, ecc.
- Ci occuperemo invece della sua **gestione**:
 - trasformazione
 - presentazione
 - diffusione
 - ...

Obiettivi del corso

- Formare figure professionali capaci di operare in ambienti digitali per la gestione e la pubblicazione di **informazioni strutturate e non strutturate** (testi e dati).
- **Non** ci occuperemo di **contenuti** in generale
 - niente film, musica, video musicali, partite di calcio, ...
- Tratteremo invece informazione:
 - **strutturata** (basi di dati, archivi, ecc.)
 - **non strutturata** (testo libero, immagini)
- **Rappresentabile graficamente**

Obiettivi del corso

- Formare figure professionali capaci di operare in ambienti digitali per la gestione e la pubblicazione di informazioni strutturate e non strutturate (testi e dati).
- Il corso avrà un approccio **tecnologico** ai problemi citati
- Studieremo standard, algoritmi, codifiche, ...
- Altri corsi si occuperanno del **contenuto** e della **forma**
 - in particolare, questo corso si coordina con **Progettazione grafica e web design**

Programma (di massima)

- **Informazione: natura, codifica, rappresentazione**
 - codifica dei testi
 - codifiche di caratteri, codifiche di documenti
 - codifica delle immagini
 - codifiche bitmap, formati, algoritmi di compressione
 - codifiche vettoriali, formati, trasformazioni
 - codifica dei colori, modelli colore, usi tipici

Programma (di massima)

- **Tecnologie di acquisizione, memorizzazione e presentazione**
 - tecnologie di scansione, lettura ottica, importazione di documenti
 - tecnologie di memorizzazione
 - tecnologie di stampa
 - tecnologie di presentazione video
 - linguaggio Postscript e formato PDF
 - formati Office proprietari (.doc) e liberi (Open Document Format)

Programma (di massima)

- **Tipografia digitale**

- storia, design e tecnologie per i font
- la formattazione del testo
 - algoritmi di formattazione di paragrafo, sillabazione, formattazione di pagina
- cenni di design e grafica
- ambienti di word processing e desktop publishing

Programma (di massima)

- **Esercitazioni**

- programmazione di algoritmi
 - transcodifica
 - formattazione
 - estrazione di informazione
- disegno di font
 - per il corpo testo (leggibilità)
 - decorativi (titolazione, “dingbats”)
 - effetti speciali

Programma (di massima)

- **Esercitazioni**

- disegno vettoriale

- creazione di un logo

- disegno bitmap / fotoritocco

- “ripulitura”, fotomontaggio

- creazione di un documento complesso

- manuale tecnico

- testo letterario

- manifesto o annuncio pubblicitario

- (valutazione delle *capacità tecniche* dimostrate)

Programma (estensione)

- Da quest'anno, il corso eroga 6 CFU anziché 5 come nel vecchio ordinamento
- Possibili estensioni del programma:
 - Editoria di testi scientifici: LaTeX
 - Editoria di testi musicali: LilyPond
 - Tecnologie per gli eBook
- Da verificare in base alla tempistica effettiva



Momento buono per ulteriori
domande e chiarimenti...

La codifica dei testi

- Un **testo** è una **sequenza di simboli**
 - Nota: un **testo** è una cosa diversa da un **documento**
- I **simboli** possono essere:
 - **caratteri** provenienti da un **alfabeto** predeterminato; hanno una rappresentazione grafica
 - **codici di controllo**, anch'essi tratti da un insieme predeterminato, solitamente non hanno rappresentazione grafica, ma alterano la rappresentazione dei caratteri

Esempio di testo

- Perché, finora cosa abbiamo proiettato??

< 'P', 'e', 'r', 'c', 'h', 'é', ',', ' ', 'f', 'i', 'n', ..., '?' >

- Nota: stiamo *scrivendo* un testo che *descrive* un testo
 - ci serve un **meta-livello** descrittivo per distinguere i caratteri **di cui** parliamo da quelli **tramite i quali** parliamo
 - qui abbiamo usato il colore

L'ambigua nozione di carattere

- Spesso si tende a identificare un **carattere** ...
 - ... con una lettera dell'alfabeto
 - ... con il segno astratto che rappresenta la lettera
 - ... con la forma concreta che tale segno assume in un certo font
 - ... con l'informazione che tale segno porta
 - ... altre opzioni?

Facciamo chiarezza!

- Un **glifo**, dal greco γλῦφω (glýphō), "incidere", è un qualunque **segno** grafico (anticamente, inciso o dipinto)
- Un **grafema** è un segno elementare, non ulteriormente divisibile, del **linguaggio scritto**
- Il grafema coincide tipicamente con una **lettera** dell'alfabeto (in senso generale) del linguaggio in questione

Facciamo chiarezza!

- Un **fonema** è invece l'unità elementare del **linguaggio parlato** (suono)
- La corrispondenza fra fonemi, grafemi, lettere e glifi può essere complessa:
 - “gn” rappresenta un solo fonema, due grafemi (digrafo), due lettere
 - “fi” consta di due grafemi, ma spesso è rappresentato da un solo glifo “**fi**” (legatura) che rappresenta due lettere e due fonemi (analoga legatura si ha per “ffi”)
 - “c/o” consta di due lettere e un simbolo, dunque tre grafemi, un solo glifo “**ç/ø**”, una lunga serie di fonemi
 - “--” è spesso usato per indicare il trattino lungo “—”; qui due glifi codificano un solo grafema che non corrisponde a nessun fonema

Facciamo chiarezza!

- Un glifo è dunque la **rappresentazione concreta** di un grafema (es., una lettera, un ideogramma cinese), di una parte di un grafema (es., un accento) o di più grafemi (es., una legatura)
- Il grafema è una unità **di testo**
- Il glifo è una unità **grafica**

Facciamo chiarezza!

- Ad ogni grafema (unità grafica astratta) possono corrispondere molteplici glifi (unità grafica concreta)
- Esempio: grafema “A”
 - A, **A**, *A*, A, **A**, *A*, **A**, **A**, *A*, *A*, **A**..
- Raccolte di glifi disegnati con lo stesso stile grafico vengono dette **font** (**fonti** o **tipi**)... ma di queste ci occuperemo più avanti

Una nozione più precisa di carattere

- I **caratteri** (simboli) di cui ci occuperemo trattando della codifica dei testi sono dunque i **grafemi**
- Infatti:
 - abbiamo caratteri che non sono lettere: %
 - abbiamo caratteri che non rappresentano fonemi: !
 - abbiamo più caratteri per la stessa lettera: a, A
 - però: grafemi multi-glifo come “--” contano come due caratteri distinti

Codifica dei testi

- Com'è noto, i calcolatori elettronici sono in grado di trattare soltanto i simboli 0 e 1 (*on e off, vero e falso,...*)
- Sequenze di 0 e 1 costituiscono *numeri binari*; con essi si possono esprimere numeri di dimensione arbitraria
- Una codifica assegna un **codice numerico** a ogni unità testuale
 - tipicamente, **un numero per carattere**

Codifica dei testi

- I codici storicamente più importanti sono:
 - codice **ASCII** (oggi di uso universale)
 - codice **EBCDIC** (oggi quasi scomparso)
- Codici moderni basati sull'ASCII
 - **ISO-8859-xxx**
 - Windows “Code Pages” (**CP-xxx**)
- Codici “del futuro”
 - **UNICODE** (e varianti di codifica)

Il codice ASCII

- **ASCII = American Standard Code for Information Interchange**
- Una codifica standard creata nel 1967 (ultima revisione del 1986)
- Basato sull'alfabeto inglese:
 - niente lettere accentate
 - niente caratteri tipici di altre lingue (ß, Ç, ñ)
 - alcuni simboli tipici dell'inglese commerciale (\$, %, @, &)

Il codice ASCII

- L'ASCII codifica:
 - **33** codici di controllo (oggi molti sono in disuso)
 - **95** caratteri
- In totale si hanno quindi 128 codici, che possono essere memorizzati in **7 bit** (cifre binarie) di informazione
- Visto che quasi tutti i computer lavorano con unità di 8 bit, un bit veniva lasciato libero per altri usi

Codici di controllo ASCII

Binario	Ottale	Decimale	Hex	Nome	Immissione da tastiera	Codice escape (C, Java)	Descrizione	
0000 0000	0	0	0	NUL	^@	\0	Null character	Carattere "nullo"
0000 0001	1	1	1	SOH	^A		Start of Header	
0000 0010	2	2	2	STX	^B		Start of Text	
0000 0011	3	3	3	ETX	^C		End of Text	
0000 0100	4	4	4	EOT	^D		End of Transmission	
0000 0101	5	5	5	ENQ	^E		Enquiry	
0000 0110	6	6	6	ACK	^F		Acknowledgment	
0000 0111	7	7	7	BEL	^G	\a	Bell	Campanella
0000 1000	10	8	8	BS	^H	\b	Backspace	Ritorno indietro di un carattere
0000 1001	11	9	9	HT	^I	\t	Horizontal Tab	Tabulazione orizzontale
0000 1010	12	10	0A	LF	^J	\n	Line feed	Nuova riga
0000 1011	13	11	0B	VT	^K		Vertical Tab	Tabulazione verticale
0000 1100	14	12	0C	FF	^L	\f	Form feed	Nuova pagina
0000 1101	15	13	0D	CR	^M	\r	Carriage return	Ritorno carrello
0000 1110	16	14	0E	SO	^N		Shift Out	
0000 1111	17	15	0F	SI	^O		Shift In	
0001 0000	20	16	10	DLE	^P		Data Link Escape	
0001 0001	21	17	11	DC1	^Q		Device Control 1 (oft. XON)	
0001 0010	22	18	12	DC2	^R		Device Control 2	
0001 0011	23	19	13	DC3	^S		Device Control 3 (oft. XOFF)	
0001 0100	24	20	14	DC4	^T		Device Control 4	
0001 0101	25	21	15	NAK	^U		Negative Acknowledgement	
0001 0110	26	22	16	SYN	^V		Synchronous Idle	
0001 0111	27	23	17	ETB	^W		End of Trans. Block	
0001 1000	30	24	18	CAN	^X		Cancel	
0001 1001	31	25	19	EM	^Y		End of Medium	
0001 1010	32	26	1A	SUB	^Z		Substitute	
0001 1011	33	27	1B	ESC	^[\e	Escape	Codice di escape
0001 1100	34	28	1C	FS	^\		File Separator	
0001 1101	35	29	1D	GS	^]		Group Separator	
0001 1110	36	30	1E	RS	^^		Record Separator	
0001 1111	37	31	1F	US	^_		Unit Separator	
0111 1111	177	127	7F	DEL	^?		Delete	Cancella (carattere successivo)

Caratteri ASCII

Simboli, numeri, punteggiatura				Lettere maiuscole				Lettere minuscole			
Binario	Decimale	Hex	Glifo	Binario	Decimale	Hex	Glifo	Binario	Decimale	Hex	Glifo
0010 0000	32	20	(spazio)	0100 0000	64	40	@	0110 0000	96	60	`
0010 0001	33	21	!	0100 0001	65	41	A	0110 0001	97	61	a
0010 0010	34	22	"	0100 0010	66	42	B	0110 0010	98	62	b
0010 0011	35	23	#	0100 0011	67	43	C	0110 0011	99	63	c
0010 0100	36	24	\$	0100 0100	68	44	D	0110 0100	100	64	d
0010 0101	37	25	%	0100 0101	69	45	E	0110 0101	101	65	e
0010 0110	38	26	&	0100 0110	70	46	F	0110 0110	102	66	f
0010 0111	39	27	'	0100 0111	71	47	G	0110 0111	103	67	g
0010 1000	40	28	(0100 1000	72	48	H	0110 1000	104	68	h
0010 1001	41	29)	0100 1001	73	49	I	0110 1001	105	69	i
0010 1010	42	2A	*	0100 1010	74	4A	J	0110 1010	106	6A	j
0010 1011	43	2B	+	0100 1011	75	4B	K	0110 1011	107	6B	k
0010 1100	44	2C	,	0100 1100	76	4C	L	0110 1100	108	6C	l
0010 1101	45	2D	-	0100 1101	77	4D	M	0110 1101	109	6D	m
0010 1110	46	2E	.	0100 1110	78	4E	N	0110 1110	110	6E	n
0010 1111	47	2F	/	0100 1111	79	4F	O	0110 1111	111	6F	o
0011 0000	48	30	0	0101 0000	80	50	P	0111 0000	112	70	p
0011 0001	49	31	1	0101 0001	81	51	Q	0111 0001	113	71	q
0011 0010	50	32	2	0101 0010	82	52	R	0111 0010	114	72	r
0011 0011	51	33	3	0101 0011	83	53	S	0111 0011	115	73	s
0011 0100	52	34	4	0101 0100	84	54	T	0111 0100	116	74	t
0011 0101	53	35	5	0101 0101	85	55	U	0111 0101	117	75	u
0011 0110	54	36	6	0101 0110	86	56	V	0111 0110	118	76	v
0011 0111	55	37	7	0101 0111	87	57	W	0111 0111	119	77	w
0011 1000	56	38	8	0101 1000	88	58	X	0111 1000	120	78	x
0011 1001	57	39	9	0101 1001	89	59	Y	0111 1001	121	79	y
0011 1010	58	3A	:	0101 1010	90	5A	Z	0111 1010	122	7A	z
0011 1011	59	3B	;	0101 1011	91	5B	[0111 1011	123	7B	{
0011 1100	60	3C	<	0101 1100	92	5C	\	0111 1100	124	7C	
0011 1101	61	3D	=	0101 1101	93	5D]	0111 1101	125	7D	}
0011 1110	62	3E	>	0101 1110	94	5E	^	0111 1110	126	7E	~
0011 1111	63	3F	?	0101 1111	95	5F	_				

Esempio di codifica ASCII

- Vediamo come viene codificata la frase “Ciao, mamma!”:

Carattere	C	i	a	o	,		m	a	m	m	a	!
Codice	67	105	97	111	44	32	109	97	109	109	97	33

- È importante ricordare che il calcolatore “vede” solo i codici numerici
- Serve un altro sistema per sapere quale codice si sta usando
 - tipicamente, è dato implicitamente
 - a volte, si specifica l'*encoding* in testa al documento

Codifica di testi in ASCII

- Un testo ASCII è codificato da una serie di caratteri, con interposti codici di controllo
- Spesso, gli unici codici di controllo usati sono quelli che indicano gli a-capo
 - ad ogni fine riga, se il testo è **preformattato** (*a-capo fisici*)
 - solo a fine paragrafo, se il testo **non è formattato** (*a-capo logici*)
 - in questo caso, il programma che visualizza il testo dovrà riformattarlo al volo

Codifica di testo in ASCII

- **Attenzione** a una trappola comune:
 - UNIX (Linux) e molti altri sistemi usano il codice **LF** (10) per indicare a-capo
 - Macintosh usa il codice **CR** (13)
 - Windows e MS-DOS usano entrambi i codici, ovvero la sequenza **CR+LF** (13+10)
- Se il testo è stato generato su un S.O. diverso e programmi usati non fanno la conversione automaticamente, può capitare di avere a che fare con linee **estremamente** lunghe!

Il codice EBCDIC

- Prodotto da IBM (tentando di competere con ASCII) nel 1963-64
- Usato su tutti i *mainframe* IBM dal System/360 in poi (tranne che da Linux su zSeries)
- Codice a 8 bit, contiene più caratteri rispetto all'ASCII, ma...
 - poco standardizzato
 - codifica bislacca (alfabeto disgiunto!)
- Interesse solo storico

Il codice EBCDIC

Hex	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
4-			â	ä	à	á	ã	å	ç	ñ	[.	<	(+	!
5-	&	é	ê	ë	è	í	î	ï	ì	ß]	\$	*)	;	^
6-	-	/	Â	Ä	À	Á	Ã	Å	Ç	Ñ	:	,	%	_	>	?
7-	ø	É	Ê	Ë	È	Í	Î	Ï	Ì	`	:	#	@	´	=	"
8-	Ø	a	b	c	d	e	f	g	h	i	«	»	ð	ý	þ	±
9-	°	j	k	l	m	n	o	p	q	r	ª	º	æ	¸	Æ	¤
A-	µ	~	s	t	u	v	w	x	y	z	ı	ı	Ð	Ÿ	Ɔ	®
B-	¢	£	¥	·	©	§	¶	¼	½	¾	¬		-	¨	´	×
C-	{	A	B	C	D	E	F	G	H	I	-	ô	ö	ò	ó	õ
D-	}	J	K	L	M	N	O	P	Q	R	¹	û	ü	ù	ú	ÿ
E-	\	÷	S	T	U	V	W	X	Y	Z	²	Ô	Ö	Ò	Ó	Õ
F-	0	1	2	3	4	5	6	7	8	9	³	Û	Ü	Ù	Ú	

CCSID500, una delle varianti di EBCDIC

Svantaggi di ASCII e EBCDIC

- ASCII, limitato a 95 caratteri, non era sufficiente per linguaggi diversi dall'Inglese (e anche per quello...)
 - Es: **perche'** invece di **perché**
- EBCDIC, mai standardizzato davvero, era inutilizzabile come formato di scambio fra macchine diverse
 - persino il codice di “a capo” poteva essere diverso su macchine diverse!

Soluzioni più moderne

- Anche i codici a 8 bit (255 caratteri) erano insufficienti a rappresentare tutti i caratteri necessari per uso generale
 - anche solo per le lingue “occidentali”!
- Soluzione: definire **una codifica diversa per ogni lingua...**
 - ISO-8859-xx
 - Windows CodePage (CP-xxx)
- ... **mantenendo la compatibilità ASCII**

Codifica ISO-8859-xx

- Sviluppata da **ISO** (International Organization for Standardization)
- Consta di 16 diverse codifiche a 8 bit, ciascuna delle quali include ASCII nei codici “bassi” e altri caratteri per un totale di 191 caratteri e 32 codici di controllo (rimangono alcuni codici liberi)
- Nota anche come ISO-Latin-xx perché codifica tutti i caratteri “latini”

ISO-8859-1

- ISO-8859-1 è stato ampiamente adottato, e può ormai essere visto come il sostituto *de facto* dello standard ASCII.
- ISO 8859-1 è adatto per i linguaggi seguenti:
 - afrikaans, basco, catalano, danese, inglese, faeroese, finlandese, francese, galiziano, irlandese, islandese, **italiano**, norvegese, olandese, portoghese, scozzese, spagnolo, svedese e tedesco.

Elenco ISO-8859-xx

Codifica	Lingue
ISO 8859-1	lingue europee occidentali (Latin-1)
ISO 8859-2	lingue europee orientali (Latin-2)
ISO 8859-3	lingue europee sudoccidentali e varie (Latin-3)
ISO 8859-4	lingue scandinave/baltiche (Latin-4)
ISO 8859-5	alfabeto romano/cirillico
ISO 8859-6	alfabeto romano/arabo
ISO 8859-7	alfabeto romano/greco
ISO 8859-8	alfabeto romano/ebraico
ISO 8859-9	latin-1 modificato per il turco (Latin-5)
ISO 8859-10	lingue lappone/nordiche/eschimesi (Latin-6)
ISO 8859-11	tailandese
ISO 8859-13	lingue baltiche (Latin-7)
ISO 8859-14	celtico (Latin-8)
ISO 8859-15	lingue europee occidentali (Latin-9)
ISO 8859-16	rumeno (Latin-10)

ISO-8859-15

- ISO-8859-15 (ISO-Latin-9) è un aggiornamento di ISO-8859-1 (Latin-1) che aggiunge:
 - il carattere € (per ovvie ragioni assente nello standard originale!)
 - alcuni caratteri rari del Francese e del Finlandese (usati anche in altre lingue)
- Rimuove inoltre alcuni codici non usati

ISO-8859-15

- Con le aggiunte, ISO-8859-15 supporta le lingue seguenti:
 - albanese, basco, bretone, catalano, danese, olandese, inglese, estone, faroese, finlandese, francese, frisone, galiziano, tedesco, islandese, groenlandese, irlandese, gaelico, italiano, latino, lussemburghese, norvegese, portoghese, reto-romanico, gaelico scozzese, spagnolo, svedese (più alcune altre che hanno una traslitterazione nell'alfabeto latino...)
- Il “nuovo ASCII”

ISO-8859-15

Decimale	Hex	Descrizione	Carattere	Decimale	Hex	Descrizione	Carattere
160	A0	NO-BREAK SPACE		208	D0	CAPITAL ICELANDIC LETTER ETH	Ð
161	A1	INVERTED EXCLAMATION MARK	¡	209	D1	CAPITAL LETTER N WITH TILDE	Ñ
162	A2	CENT SIGN	¢	210	D2	CAPITAL LETTER O WITH GRAVE ACCENT	Ò
163	A3	POUND SIGN	£	211	D3	CAPITAL LETTER O WITH ACUTE ACCENT	Ó
164	A4	EURO SIGN	€	212	D4	CAPITAL LETTER O WITH CIRCUMFLEX ACCENT	Ô
165	A5	YEN SIGN	¥	213	D5	CAPITAL LETTER O WITH TILDE	Õ
166	A6	CAPITAL LETTER S WITH CARON	Š	214	D6	CAPITAL LETTER O WITH DIAERESIS	Ö
167	A7	PARAGRAPH SIGN	§	215	D7	MULTIPLICATION SIGN	×
168	A8	SMALL LETTER S WITH CARON	š	216	D8	CAPITAL LETTER O WITH OBLIQUE STROKE	Ø
169	A9	COPYRIGHT SIGN	©	217	D9	CAPITAL LETTER U WITH GRAVE ACCENT	Ù
170	AA	FEMININE ORDINAL INDICATOR	ª	218	DA	CAPITAL LETTER U WITH ACUTE ACCENT	Ú
171	AB	LEFT ANGLE QUOTATION MARK	«	219	DB	CAPITAL LETTER U WITH CIRCUMFLEX ACCENT	Û
172	AC	NOT SIGN	¬	220	DC	CAPITAL LETTER U WITH DIAERESIS	Ü
173	AD	SOFT HYPHEN	-	221	DD	CAPITAL LETTER Y WITH ACUTE ACCENT	Ý
174	AE	REGISTERED TRADE MARK SIGN	®	222	DE	CAPITAL ICELANDIC LETTER THORN	Þ
175	AF	MACRON	-	223	DF	SMALL GERMAN LETTER SHARP s	ß
176	B0	DEGREE SIGN, RING ABOVE	°	224	E0	SMALL LETTER a WITH GRAVE ACCENT	à
177	B1	PLUS-MINUS SIGN	±	225	E1	SMALL LETTER a WITH ACUTE ACCENT	á
178	B2	SUPERSCRIPIT TWO	²	226	E2	SMALL LETTER a WITH CIRCUMFLEX ACCENT	â
179	B3	SUPERSCRIPIT THREE	³	227	E3	SMALL LETTER a WITH TILDE	ã
180	B4	CAPITAL LETTER Z WITH CARON	Ž	228	E4	SMALL LETTER a WITH DIAERESIS	ä
181	B5	MICRO SIGN	μ	229	E5	SMALL LETTER a WITH RING ABOVE	å
182	B6	PILCROW SIGN	¶	230	E6	SMALL DIPHTHONG a WITH e	æ
183	B7	MIDDLE DOT	·	231	E7	SMALL LETTER c WITH CEDILLA	ç
184	B8	SMALL LETTER Z WITH CARON	ž	232	E8	SMALL LETTER e WITH GRAVE ACCENT	è
185	B9	SUPERSCRIPIT ONE	¹	233	E9	SMALL LETTER e WITH ACUTE ACCENT	é
186	BA	MASCULINE ORDINAL INDICATOR	º	234	EA	SMALL LETTER e WITH CIRCUMFLEX ACCENT	ê
187	BB	RIGHT ANGLE QUOTATION MARK	»	235	EB	SMALL LETTER e WITH DIAERESIS	ë
188	BC	CAPITAL OE DIGRAPH	Œ	236	EC	SMALL LETTER i WITH GRAVE ACCENT	ì
189	BD	SMALL OE DIGRAPH	œ	237	ED	SMALL LETTER i WITH ACUTE ACCENT	í
190	BE	CAPITAL LETTER Y WITH DIAERESIS	ÿ	238	EE	SMALL LETTER i WITH CIRCUMFLEX ACCENT	î
191	BF	INVERTED QUESTION MARK	¿	239	EF	SMALL LETTER i WITH DIAERESIS	ï
192	C0	CAPITAL LETTER A WITH GRAVE ACCENT	À	240	F0	SMALL ICELANDIC LETTER eth	ð
193	C1	CAPITAL LETTER A WITH ACUTE ACCENT	Á	241	F1	SMALL LETTER n WITH TILDE	ñ
194	C2	CAPITAL LETTER A WITH CIRCUMFLEX ACCENT	Â	242	F2	SMALL LETTER o WITH GRAVE ACCENT	ò
195	C3	CAPITAL LETTER A WITH TILDE	Ã	243	F3	SMALL LETTER o WITH ACUTE ACCENT	ó
196	C4	CAPITAL LETTER A WITH DIAERESIS	Ä	244	F4	SMALL LETTER o WITH CIRCUMFLEX ACCENT	ô
197	C5	CAPITAL LETTER A WITH RING ABOVE	Å	245	F5	SMALL LETTER o WITH TILDE	õ
198	C6	CAPITAL DIPHTHONG A WITH E	Æ	246	F6	SMALL LETTER o WITH DIAERESIS	ö
199	C7	CAPITAL LETTER C WITH CEDILLA	Ç	247	F7	DIVISION SIGN	÷
200	C8	CAPITAL LETTER E WITH GRAVE ACCENT	È	248	F8	SMALL LETTER o WITH OBLIQUE STROKE	ø
201	C9	CAPITAL LETTER E WITH ACUTE ACCENT	É	249	F9	SMALL LETTER u WITH GRAVE ACCENT	ù
202	CA	CAPITAL LETTER E WITH CIRCUMFLEX ACCENT	Ê	250	FA	SMALL LETTER u WITH ACUTE ACCENT	ú
203	CB	CAPITAL LETTER E WITH DIAERESIS	Ë	251	FB	SMALL LETTER u WITH CIRCUMFLEX ACCENT	û
204	CC	CAPITAL LETTER I WITH GRAVE ACCENT	Ì	252	FC	SMALL LETTER u WITH DIAERESIS	ü
205	CD	CAPITAL LETTER I WITH ACUTE ACCENT	Í	253	FD	SMALL LETTER y WITH ACUTE ACCENT	ý
206	CE	CAPITAL LETTER I WITH CIRCUMFLEX ACCENT	Î	254	FE	SMALL ICELANDIC LETTER THORN	þ

CodePage-xx

- Idea simile a quella di ISO-8859-xx, ma
 - sottilmente incompatibile
 - non standardizzata
- Codifica introdotta da IBM e Microsoft
 - ciascuno dei due ha poi continuato a “inventare” nuove codepage secondo necessità
 - Usata nei mainframe IBM, in PC-DOS, MS-DOS, Windows

Alcuni esempi

- Codepage definite da IBM per il PC-DOS (e MS-DOS):

Codepage	Uso
CP437	Codifica usata nel PC IBM originale (1981)
CP737	Codifica MS-DOS per l'alfabeto greco
CP850	Codifica dell'alfabeto latino e greco + simboli grafici
CP852	Codifica per le lingue del centro europa (polacco, romeno, ceco, slovacco)
CP855	Codifica per l'alfabeto cirillico
CP857	Codifica per l'alfabeto turco
CP858	CP850 + simbolo euro (al posto della i senza puntino)
CP860	Codifica per il Portoghese
CP861	Codifica per l'Islandese e altri linguaggi nordici
CP863	Codifica estesa per il Francese (usato in Canada)
CP865	Codifica per il Danese e il Norvegese (e altri linguaggi nordici escluso l'Islandese)
CP866	Altra codifica per il cirillico
CP869	Altra codifica per il greco

Alcuni esempi

- Codepage definite da Microsoft per Windows (anche note come Windows-xx)

Codepage	Uso
-----------------	------------

CP1250	Codifica per le lingue del centro Europa (polacco, romeno, slovacco, ungherese, ...)
--------	--

CP1251	Codifica per l'alfabeto cirillico
--------	-----------------------------------

CP1252	Codifica per le lingue dell'Europa occidentale (incluso l'Italiano)
--------	---

CP1253	Codifica per l'alfabeto greco
--------	-------------------------------

CP1254	Codifica per l'alfabeto turco
--------	-------------------------------

CP1255	Codifica per l'alfabeto ebraico
--------	---------------------------------

CP1256	Codifica per l'alfabeto arabo
--------	-------------------------------

CP1257	Codifica per le lingue baltiche
--------	---------------------------------

CP1258	Codifica per il Vietnamita
--------	----------------------------

...

Alcuni esempi

- Sulla scia di ISO e CP, altri costruttori e paesi hanno introdotto proprie codifiche “ad hoc”:
 - GB2312 per il cinese
 - KOI-8 per il russo e il bulgaro, KOI-8-U per l'Ucraino
 - Apple con il Macintosh introdusse alcune codifiche proprie
 - ancora una volta, leggermente incompatibili con tutte le altre...

CP437

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
0.	NULL 0 0	☉ 263A 1	☽ 263B 2	♥ 2665 3	♦ 2666 4	♣ 2663 5	♠ 2660 6	• 2022 7	♠ 25D8 8	○ 25CB 9	● 25D9 10	♂ 2642 11	♀ 2640 12	♪ 266A 13	♪ 266B 14	* 263C 15
1.	▶ 25BA 16	◀ 25C4 17	↕ 2195 18	!! 203C 19	¶ B6 20	§ A7 21	■ 25AC 22	‡ 21A8 23	↑ 2191 24	↓ 2193 25	→ 2192 26	← 2190 27	↳ 221F 28	↔ 2194 29	▲ 25B2 30	▼ 25BC 31
2.	20 32 0	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3.	30 48	1 49	2 50	3 51	4 52	5 53	6 54	7 55	8 56	9 57	:	;	<	=	>	?
4.	@ 40 64	A 41 65	B 42 66	C 43 67	D 44 68	E 45 69	F 46 70	G 47 71	H 48 72	I 49 73	J 4A 74	K 4B 75	L 4C 76	M 4D 77	N 4E 78	O 4F 79
5.	P 50 80	Q 51 81	R 52 82	S 53 83	T 54 84	U 55 85	V 56 86	W 57 87	X 58 88	Y 59 89	Z 5A 90	[5B 91	\ 5C 92] 5D 93	^ 5E 94	_ 5F 95
6.	` 60 96	a 61 97	b 62 98	c 63 99	d 64 100	e 65 101	f 66 102	g 67 103	h 68 104	i 69 105	j 6A 106	k 6B 107	l 6C 108	m 6D 109	n 6E 110	o 6F 111
7.	p 70 112	q 71 113	r 72 114	s 73 115	t 74 116	u 75 117	v 76 118	w 77 119	x 78 120	y 79 121	z 7A 122	{ 7B 123	 7C 124	}	~ 7E 126	△ 2302 127
8.	Ç C7 128	Û FC 129	É E9 130	Â E2 131	Ä E4 132	À E0 133	Å E5 134	Ç E7 135	Ê EA 136	Ë EB 137	È E8 138	Ï EF 139	Î EE 140	Ì EC 141	Ä C4 142	Å C5 143
9.	É C9 144	æ E6 145	Æ C6 146	ô F4 147	ö F6 148	ò F2 149	û FB 150	ù F9 151	ÿ FF 152	Ö D6 153	Ü DC 154	ç A2 155	£ A3 156	¥ A5 157	Pts 20A7 158	f 192 159
A.	á E1 160	í ED 161	ó F3 162	ú FA 163	ñ F1 164	Ñ D1 165	ª AA 166	º BA 167	¿ BF 168	¬ 2310 169	¬ AC 170	½ BD 171	¼ BC 172	i A1 173	« AB 174	» BB 175
B.	⌘ 2591 176	⌘ 2592 177	⌘ 2593 178	 2502 179	 2524 180	 2561 181	 2562 182	⌘ 2556 183	⌘ 2555 184	⌘ 2563 185	⌘ 2551 186	⌘ 2557 187	⌘ 255D 188	⌘ 255C 189	⌘ 255B 190	⌘ 2510 191
C.	⌘ 2514 192	⌘ 2534 193	⌘ 252C 194	 251C 195	- 2500 196	+ 253C 197	 255E 198	 255F 199	⌘ 255A 200	⌘ 2554 201	⌘ 2569 202	⌘ 2566 203	⌘ 2560 204	= 2550 205	⌘ 256C 206	⌘ 2567 207
D.	⌘ 2568 208	⌘ 2564 209	⌘ 2565 210	⌘ 2559 211	⌘ 2558 212	⌘ 2552 213	⌘ 2553 214	⌘ 256B 215	⌘ 256A 216	⌘ 2518 217	⌘ 250C 218	⌘ 2588 219	⌘ 2584 220	⌘ 258C 221	⌘ 2590 222	⌘ 2580 223
E.	α 3B1 224	β DF 225	Γ 393 226	π 3C0 227	Σ 3A3 228	σ 3C3 229	μ B5 230	τ 3C4 231	Φ 3A6 232	Θ 398 233	Ω 3A9 234	δ 3B4 235	∞ 221E 236	φ 3C6 237	ε 3B5 238	η 2229 239
F.	≡ 2261 240	± B1 241	≥ 2265 242	≤ 2264 243	∫ 2320 244	∫ 2321 245	÷ F7 246	≈ 2248 247	° B0 248	· 2219 249	· B7 250	√ 221A 251	ⁿ 207F 252	² B2 253	■ 25A0 254	A0 255

Il problema con le Codepage

- Spesso la definizione segue criteri più commerciali che linguistici
 - La CP850 venne definita in quel modo (faccine, notine, semi delle carte da gioco), per dimostrare che il PC poteva far girare un programma simile a una nota macchina di word processing Wang (!)
 - CP diverse sono state definite per paesi con alfabeti simili, ma economie diverse
 - versioni del S.O. per il paese più ricco venivano fatte pagare più care rispetto ai “vicini poveri”

Il problema con le Codepage

- A volte CP diverse sono del tutto identiche, a parte uno o due caratteri
 - È estremamente difficile riconoscere a prima vista quale CP è stata usata per codificare un certo testo
- Più CP possono essere usate per la stessa lingua
 - incompatibilità anche all'interno dello stesso paese e fra testi prodotti sulla stessa macchina

Problemi comuni a ISO e CP

- È possibile codificare al più 256 caratteri distinti
 - mantenendo la compatibilità ASCII, ancora meno!
 - lingue ideografiche (cinese, giapponese, coreano) o complesse sono escluse
- Lo **stesso codice** numerico può corrispondere a **caratteri diversi** in codifiche diverse
- Lo **stesso carattere** può corrispondere a **codici diversi** in codifiche diverse

Approfondimenti

- Il sito <http://homepages.cwi.nl/~dik/english/codes/stand.html> contiene un elenco di codifiche “storiche” e delle loro successive versioni
- Il sito <http://www.iana.org/assignments/character-sets> elenca i 250 set di caratteri ufficialmente riconosciuti da IANA (Internet Assigned Numbers Authority)
- Il sito <http://www.i18nguy.com/unicode/codepages.html> contiene molte tabelle di codifiche e link utili sulle varie codepage e altre codifiche
- Su Linux, i comandi “man ascii” e “man latin1” mostrano le rispettive tabelle (analogamente per altre codifiche)