


Unicode

La codifica dei testi antichi

Monica L'Erario
lerario@uniroma3.it
L.I.S.A.
Università Roma Tre

Alfabeti a 7 bit (ASCII)

Lo standard ISO 646

ASCII (American Standard Code for Information Interchange)
standard americano a 128 caratteri:
il set grafico è composto da **94**
caratteri dei quali **13**, nella versione
ISO 646 sono riservati alle varianti
nazionali

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

Alfabeti a 8 bit (ASCII)

Lo standard ISO 8859

lo scopo dello standard ISO 8859 è quello di aggiungere altri caratteri, quali lettere accentate, particolari segni di interpunzione, lettere non latine presenti nelle diverse lingue europee. la seconda parte della tabella caratterizza, NB utilizzando sempre gli stessi codici, le varianti nazionali dello standard ISO.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	DLE	SP	0	@	P	`	p			NBSP	°	À	Đ	à	ö
1	SOH	DC1	!	1	A	Q	a	q			ı	±	Á	Ñ	á	ñ
2	STX	DC2	"	2	B	R	b	r			¢	²	Â	Ò	â	ò
3	ETX	DC3	#	3	C	S	c	s			£	³	Ã	Ó	ã	ó
4	EOT	DC4	\$	4	D	T	d	t			¤	´	Ä	Ô	ä	ô
5	ENQ	NAK	%	5	E	U	e	u			¥	µ	Å	Ö	å	ö
6	ACK	SYN	&	6	F	V	f	v			¦	¶	Æ	Ö	æ	ö
7	BEL	ETB	'	/	G	W	g	w			§	·	Ç	×	ç	÷
8	BS	CAN	(8	H	X	h	x			¨	,	È	Ø	è	ø
9	HT	EM)	9	I	Y	i	y			©	¹	É	Ù	é	ù
A	LF	SUB	*	:	J	Z	j	z			ª	º	Ê	Ú	ê	ú
B	VT	ESC	+	;	K	[k	{			«	»	Ë	Û	ë	û
C	FF	FS	,	<	L	\	l				¬	¼	Ì	Ü	ì	ü
D	CR	GS	-	=	M]	m	}			-	½	Í	Ý	í	ý
E	SO	RS	.	>	N	^	n	~			®	¾	Î	Þ	î	þ
F	SI	US	/	?	O	_	o	DEL			¯	¿	Ï	ß	ï	ÿ

Lo standard ISO 8859

Le varianti nazionali

Charset ISO-8859-1 (Latin 1)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0			0	@	P	·	p				°	À	D	à	ð	
1		!	1	A	Q	a	q			¡	±	Á	Ñ	á	ñ	
2		"	2	B	R	b	r			¢	²	Â	Ò	â	ò	
3		#	3	C	S	c	s			£	³	Ã	Ó	ã	ó	
4		\$	4	D	T	d	t			¤	´	Ä	Ö	ä	ö	
5		%	5	E	U	e	u			¥	µ	Å	Õ	å	õ	
6		&	6	F	V	f	v			¦	¶	Æ	Ö	æ	ö	
7		'	7	G	W	g	w			§	·	Ç	×	ç	÷	
8		(8	H	X	h	x			¨	¸	È	Ø	è	ø	
9)	9	I	Y	i	y			©	¹	É	Ù	é	ù	
A		*	:	J	Z	j	z			ª	º	Ê	Û	ê	û	
B		+	;	K	[k	{			«	»	Ë	Ü	ë	ü	
C		,	<	L	\	l				¬	¼	Ï	Ü	ï	ü	
D		-	=	M]	m	}			–	½	Í	Ý	í	ý	
E		.	>	N	^	n	~			®	¾	Î	Þ	î	þ	
F		/	?	O	_	o	~			–	¿	Ï	ß	ï	ÿ	

www.langbox.com - LangBox International

Charset ISO 8859-7 (Greek)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0			0	@	Γ	`	ρ				°	İ	Π	Ù	π	
1		!	1	A	Q	a	q			¡	±	Α	Ρ	α	ρ	
2		"	2	B	R	b	r			¢	²	Β		β	ç	
3		#	3	C	S	c	s			£	³	Γ	Σ	γ	σ	
4		\$	4	D	T	d	t			¤	´	Δ	Τ	δ	τ	
5		%	5	E	U	e	u			¥	µ	Ε	Υ	ε	υ	
6		&	6	F	V	f	v			¦	¶	Ζ	Φ	ζ	φ	
7		'	7	G	W	g	w			§	·	Η	Χ	η	χ	
8		(8	H	X	h	x			¨	¸	Θ	Ψ	θ	ψ	
9)	9	I	Y	i	y			©	¹	Η	Ι	ι	ο	
A		~	:	J	Z	j	z			ª	º	Κ	Ϊ	κ	ϊ	
B		+	;	K	[k	{			«	»	Λ	Ϋ	λ	ϋ	
C		,	<	L	\	l				¬	¼	Μ	Ϝ	μ	ϝ	
D		-	=	M]	m	}			–	½	Ν	Ϛ	ν	ϛ	
E		.	>	N	^	n	~			®	¾	Ξ	Ϟ	ξ	ϟ	
F		/	?	O	_	o	~			–	¿	Ο	Ϡ	ο	ϡ	

www.langbox.com - LangBox International



Lo standard ISO 10646

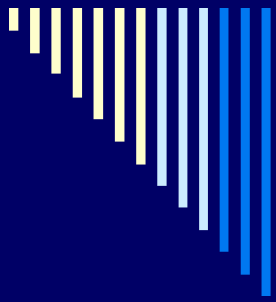
La codifica a 8 bit non è più sufficiente

Trasportabilità del formato dei documenti

- da un computer ad un altro (anche in diversi paesi)
- da un sistema operativo ad un altro
- su differenti browsers

Multilinguismo del contenuto dei documenti

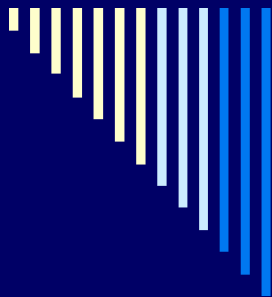
Universal Character Set



Unicode

www.unicode.org

Codifica a 16 bit: $256 * 256$ combinazioni
= 65.535 caratteri



Unicode e gli alfabeti antichi

Particolare attenzione è stata posta, sin dall'inizio del progetto Unicode, alla codifica degli alfabeti antichi, e quindi alle esigenze degli studiosi. Unicode è un progetto in divenire, è arrivato alla versione 4.0.1, nella quale è stata inserita una sottotabella, meglio code chart, per il copto.

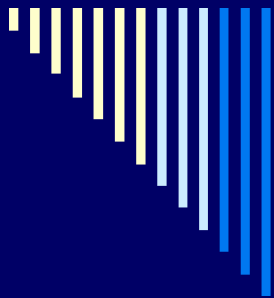
Nella prossima release dovrebbero essere inserite le posizioni per la scrittura cuneiforme e per quella fenicia (5.0).

The Unicode Character Code Charts

SCRIPTS | SYMBOLS AND PUNCTUATION | INDEX | CONVENTIONS AND RELATED LINKS

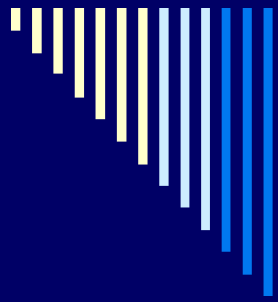
European Alphabets (see also Comb. Marks)	African Scripts	Indic Scripts	East Asian Scripts	Central Asian Scripts
Armenian	Ethiopic	Bengali	Han Ideographs	Kharoshthi
Armenian	Ethiopic Supplement	Devanagari	Unified CJK Ideographs (5MB)	Mongolian
<i>Armenian Ligatures</i>	Ethiopic Extended	Gujarati	CJK Ideographs Ext. A (2MB)	Phags-Pa (5.0)
Coptic	Other	Gurmukhi	CJK Ideographs Ext. B (13MB)	Tibetan
Coptic	N'Ko (5.0)	Kannada	Compatibility Ideographs (.5MB)	
<i>Coptic in Greek block</i>	Tifinagh	Limbu	Compatibility Ideo. Suppl. (.5MB)	
Cyrillic	Middle Eastern Scripts	Malayalam	Kanbun	
Cyrillic	Arabic	Oriya	(see also Unihan Database)	Ancient Scripts
Cyrillic Supplement	Arabic	Sinhala	Radicals and Strokes	Ancient Greek
Georgian	Arabic Supplement	Syloti Nagri	CJK Radicals	Ancient Greek Numbers
Georgian	Arabic Presentation Forms A	Tamil	KangXi Radicals	Ancient Greek Musical
<i>Georgian Supplement</i>	Arabic Presentation Forms B	Telugu	CJK Strokes	Cuneiform
Greek	Hebrew	Philippine Scripts	Ideographic Description	Cuneiform (5.0)
Greek	Hebrew	Buhid	Chinese-specific	Cuneiform Numbers (5.0)
Greek Extended	<i>Hebrew Presentation Forms</i>	Hanunoo	Bopomofo	Old Persian
(see also Ancient Greek)	Other ME Scripts	Tagalog	Japanese-specific	Ugaritic
Latin	Syriac	Tagbanwa	Hiragana	Linear B
Basic Latin	Thaana		Katakana	Linear B Syllabary
Latin-1	American scripts	South East Asian	Katakana Phonetic Ext.	Linear B Ideograms
Latin Extended A	Canadian Syllabics	Buginese	<i>Halfwidth Katakana</i>	Other Ancient Scripts
Latin Extended B	Cherokee	Balinese (5.0)	Korean-specific	Aegean Numbers
Latin Extended C (5.0)	Deseret	Khmer	Hangul Syllables (7MB)	Counting Rod Num. (5.0)
Latin Extended Additional		Lao	Hangul Jamo	Cypriot Syllabary
<i>Latin Ligatures</i>	Other Scripts	Myanmar	Hangul Compatibility Jamo	Gothic
<i>Fullwidth Latin Letters</i>	Shavian	New Tai Lue	<i>Halfwidth Jamo</i>	Old Italic
Small Forms	Osmanya	Tai Le	Yi	Ogham
(see also Phonetic Symbols)	Glagolitic	Thai	Yi (.6MB)	Runic
			Yi Radicals	Phoenician (5.0)

To access a chart, click on its entry in the table. The charts are [PDF](#) files, and some of them may be very large. For frequent access to the same chart, right-click and save the file to your disk. For an **alphabetical** index of characters and blocks, use the [Unicode Character Names Index](#). For terms of use, other information and [additional](#) ways to access the code charts, see [About the Code Charts](#).



Redigere documenti digitali utilizzando la codifica Unicode. I fonts (font = veste tipografica di un set di caratteri) normalmente utilizzati fino ad oggi si basano sulla codifica ASCII e sullo standard ISO 8859. Quindi su set di 256 caratteri al massimo.

Ma adesso le possibilità di Unicode raggiungono e superano il **milione** di segni grafici: *si tratta non più di tanti alfabeti locali ma di un solo grandissimo alfabeto multilingue.*



Unicode e gli alfabeti antichi

**Tabella caratteri greci +
14 caratteri copti derivanti
dal demotico**

Unicode e gli alfabeti antichi



Tabella caratteri greci con
segni diacritici

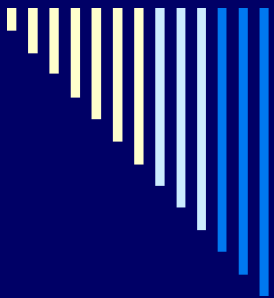
Si tratta di 'combinig
characters' o caratteri
composti



Unicode e gli alfabeti antichi

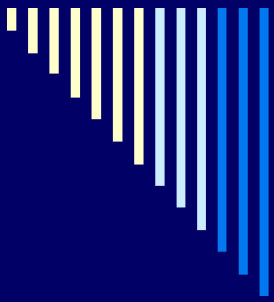
Tabelle relative:

- 1. Numeri acrofonici greci (Attica, Tespi, Troezene, Epidauro Naxos, Delfi etc.) e numerali in lettere greche**
 - 2. Notazioni musicali greche vocaliche e strumentali)**
-



Unicode e gli alfabeti antichi

**Tabella relativa
all'alfabeto persiano**



Unicode e gli alfabeti antichi

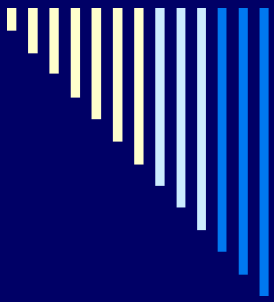
**Tabella relativa
all'alfabeto di Ugarit**



Unicode e gli alfabeti antichi

Tabelle relative a:

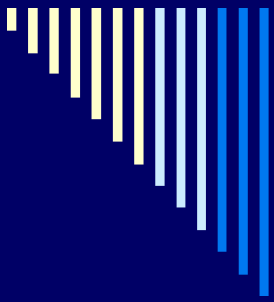
- 1. Pittogrammi della Lineare B**
- 2. Lineare B sillabica**



Unicode e gli alfabeti antichi

**Tabelle relative a
numerali e unità
di misura egei**

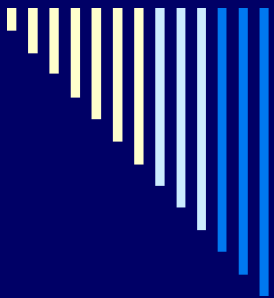




Unicode e gli alfabeti antichi

**Tabelle relative a
scrittura sillabica
cipriota**





Unicode e gli alfabeti antichi

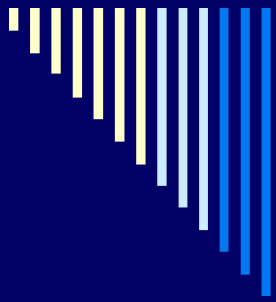
**Tabelle relative a
italico antico**





Unicode e gli alfabeti antichi

Versione 4.0.1
Nuova tabella per
l'alfabeto copto.
N.B. le 14 lettere
derivate dal
demotico
rimangono nella
tabella 'greek and
coptic'



Codifica



Font

Un font comprensivo di tutti i caratteri di tutte le lingue non è ancora stato creato, mi sembra, ma avrà un problema di dimensioni: sarà enorme. Arial Unicode MS di Microsoft è al momento il font che copre il più ampio spettro di caratteri (contiene 51.180 caratteri nella versione 0.86*): pesa 22Mb...times new roman ne pesa complessivamente 1,4 c.ca, Impact solo 133Kb.

*fonte: <http://www.alanwood.net/unicode>



Come procedere?

Mi sembra che le strade siano fundamentalmente due:

- la creazione e la successiva installazione sui computer di un solo grande font su base Unicode che permetta di scrivere e ricevere o semplicemente consultare su web in tutte le lingue
- la creazione di fonts su base unicode 'dedicati' a particolari categorie di utenti

Immaginando che:

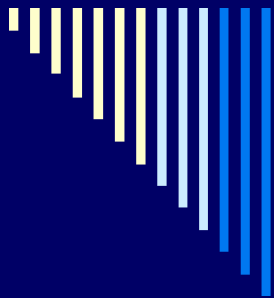
il manager turco che ha contatti e partnerships con la Cina sarà interessato ad un font che gli permetta di redigere testi nelle due lingue, e di leggere agevolmente documenti digitali inviati dai suoi partners, e viceversa.

O ancora...

Lo studioso di dialettologia greca avrà necessità di avere un font che gli permetta: di redigere testi in greco antico e nella propria lingua e magari nella lingua del paese che ospita il prossimo convegno al quale partecipa

Il linguista che sta lavorando ad un vocabolario arabo-russo troverà il font che comprende i set di caratteri arabi e russi...e così via.

A me pare che questa sia, al momento, la soluzione migliore, che permette anche di salvare l'aspetto 'estetico'- tipografico dei fonts.



Pace
εἰρήνη

שלום

سلا

Hasîti

□□□□□□

Bariş

和平

Мир

Monica L'Erario
L.I.S.A.

Laboratorio Informatico per gli Studi di Antichità
Università Roma Tre

lerario@uniroma3.it

http://host.uniroma3.it/dipartimenti/mondo_antico/LISA.htm