

Introduction to Unicode



By:
Atif Gulzar
Center for Research in Urdu
Language Processing

Introduction to Unicode

■ Unicode

- Why Unicode ?
- What is Unicode ?
- Unicode Architecture

Why Unicode?

Pre-Unicode Standards and their Limitations

- ASCII by ANSI in 1964 (7 bit code)
- ISO adopt ASCII in 1967 as ISO 646
- ISO 2022 (8 bit code)
- ISO 8859
 - ISO 8859 is a family of 16 Standards
- Code Page
- And plenty of standards for East Asian languages

ISO 8859 cont.

- ISO 8859-1, Latin-1, Western European
- ISO 8859-2, Latin-2, Eastern European
- ISO 8859-3, Latin-3, Southern European
- ISO 8859-4, Latin-4, Northern European
- ISO 8859-5, Cyrillic, Russian, Bulgarian..
- ISO 8859-6, Arabic, Arabic
- ISO 8859-7, Greek, Greek
- ISO 8859-8, Hebrew, Hebrew

ISO 8859

- ISO 8859-9, Latin-5, Turkish
- ISO 8859-10, Latin-6, Northern European
- ISO 8859-11, Thai, Thai
- ISO 8859-13, Latin-7, Baltic
- ISO 8859-14, Latin-8, Celtic
- ISO 8859-15, Latin-9, Western European
- ISO 8859-16, Latin-10, Eastern European

ASCII

Upper case (A-Z)	26
Digits (0-9)	10
Space	1
Punctuation marks (.,+{)%)	32
Lower case (a-z)	26
Control characters (tab, cr, lf)	33
=====	
Total	128

ASCII

code page

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL 0000	STX 0001	SOT 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
10	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
20	SP 0020	! 0021	" 0022	# 0023	\$ 0024	% 0025	& 0026	' 0027	(0028) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
30	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
40	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
50	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C] 005D	^ 005E	_ 005F
60	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
70	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	 007C	} 007D	~ 007E	DEL 007F

ANSI 1252 code page

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL 0000	STX 0001	SOT 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
10	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
20	SP 0020	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
30	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	:	;	<	=	>	?
40	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
50	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C] 005D	^ 005E	_ 005F
60	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
70	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	 007C	}	~ 007E	DEL 007F
80	€ 20AC		ƒ 201A	„ 0192	„ 201E	… 2026	† 2020	‡ 2021	^ 02C6	‰ 2030	Š 0160	< 2039	Œ 0152		Ž 017D	
90		˘ 2018	˙ 2019	˚ 201C	˛ 201D	• 2022	– 2013	— 2014	ˆ 02DC	™ 2122	š 0161	> 203A	œ 0153		ž 017E	ÿ 0178
A0	NBSP 00A0	ı 00A1	ç 00A2	£ 00A3	¤ 00A4	¥ 00A5	¦ 00A6	§ 00A7	¨ 00A8	© 00A9	ª 00AA	« 00AB	¬ 00AC	– 00AD	® 00AE	¯ 00AF
B0	° 00B0	± 00B1	² 00B2	³ 00B3	´ 00B4	µ 00B5	¶ 00B6	· 00B7	¸ 00B8	¹ 00B9	º 00BA	» 00BB	¼ 00BC	½ 00BD	¾ 00BE	¿ 00BF
C0	À 00C0	Á 00C1	Â 00C2	Ã 00C3	Ä 00C4	Å 00C5	Æ 00C6	Ç 00C7	È 00C8	É 00C9	Ê 00CA	Ë 00CB	Ì 00CC	Í 00CD	Î 00CE	Ï 00CF
D0	Ð 00D0	Ñ 00D1	Ò 00D2	Ó 00D3	Ô 00D4	Õ 00D5	Ö 00D6	× 00D7	Ø 00D8	Ù 00D9	Ú 00DA	Û 00DB	Ü 00DC	Ý 00DD	Þ 00DE	ß 00DF
E0	à 00E0	á 00E1	â 00E2	ã 00E3	ä 00E4	å 00E5	æ 00E6	ç 00E7	è 00E8	é 00E9	ê 00EA	ë 00EB	ì 00EC	í 00ED	î 00EE	ï 00EF
F0	ø 00F0	ñ 00F1	ò 00F2	ó 00F3	ô 00F4	õ 00F5	ö 00F6	÷ 00F7	ø 00F8	ù 00F9	ú 00FA	û 00FB	ü 00FC	ý 00FD	þ 00FE	ÿ 00FF

ANSI 1256 code page

Arabic

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL 0000	STX 0001	SOT 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
10	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
20	SP 0020	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
30	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	:	;	<	=	>	?
40	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
50	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C] 005D	^ 005E	_ 005F
60	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
70	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	 007C	} 007D	~ 007E	DEL 007F
80	€ 20AC	پ 067E	، 201A	ف 0192	// 201E	... 2026	† 2020	‡ 2021	^ 02C6	§ 2030	< 2039	€ 0152	ج 0686	ج 0636		
90	گ 06AF	، 2018	، 2019	“ 201C	” 201D	▪ 2022	- 2013	- 2014	™ 2122		> 203A	œ 0153		200C	200D	
A0	NBSP 00A0	، 060C	¢ 00A2	£ 00A3	¤ 00A4	¥ 00A5	¦ 00A6	§ 00A7	¨ 00A8	© 00A9		« 00AB	¬ 00AC	- 00AD	® 00AE	¯ 00AF
B0	° 00B0	± 00B1	² 00B2	³ 00B3	´ 00B4	µ 00B5	¶ 00B6	· 00B7	¸ 00B8	¹ 00B9	º 061B	» 00BB	¼ 00BC	½ 00BD	¾ 00BE	¿ 061F
C0		ء 0621	آ 0622	أ 0623	ؤ 0624	إ 0625	ئ 0626	ا 0627	ب 0628	ب 0629	ت 062A	ث 062B	ج 062C	ح 062D	خ 062E	د 062F
D0	ذ 0630	ر 0631	ز 0632	س 0633	ش 0634	ص 0635	ض 0636	× 00D7	ط 0637	ظ 0638	ع 0639	غ 063A	- 0640	ف 0641	ق 0642	ك 0643
E0	à 00E0	آ 0644	â 00E2	م 0645	ن 0646	و 0647	و 0648	ç 00E7	è 00E8	é 00E9	ê 00EA	ë 00EB	ى 0649	ي 064A	î 00EE	ï 00EF
F0	، 064B	، 064C	، 064D	، 064E	ô 00F4	، 064F	، 0650	÷ 00F7	، 0651	ù 00F9	، 0652	û 00FB	ü 00FC	200E	200F	

ANSI 1252 code page

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL 0000	STX 0001	SOT 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
10	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
20	SP 0020	! 0021	" 0022	# 0023	\$ 0024	% 0025	& 0026	' 0027	(0028) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
30	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
40	Ø 0040	À 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
50	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C] 005D	^ 005E	_ 005F
60	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
70	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	 007C	} 007D	~ 007E	DEL 007F
80	€ 20AC		ƒ 201A	„ 0192	… 201E	† 2026	‡ 2020	^ 02C6						Ž 017D		
90		˘ 2018	˙ 2019	˚ 201C	˛ 201D	• 2022	– 2013	— 2014	˝ 02DC					ž 017E	ÿ 0178	
A0	NBSP 00A0	ı 00A1	ç 00A2	£ 00A3	¤ 00A4	¥ 00A5	ı 00A6	§ 00A7	¨ 00A8	ˆ 00A9	˜ 00AA	˘ 00AB	˙ 00AC	˚ 00AD	˛ 00AE	˜ 00AF
B0	° 00B0	± 00B1	² 00B2	³ 00B3	´ 00B4	µ 00B5	¶ 00B6	· 00B7	¸ 00B8	¹ 00B9	º 00BA	» 00BB	¼ 00BC	½ 00BD	¾ 00BE	¿ 00BF
C0	À 00C0	Á 00C1	Â 00C2	Ã 00C3	Ä 00C4	Å 00C5	Æ 00C6	Ç 00C7	È 00C8	É 00C9	Ê 00CA	Ë 00CB	Ì 00CC	Í 00CD	Î 00CE	Ï 00CF
D0	Ð 00D0	Ñ 00D1	Ò 00D2	Ó 00D3	Ô 00D4	Õ 00D5	Ö 00D6	× 00D7	Ø 00D8	Ù 00D9	Ú 00DA	Û 00DB	Ü 00DC	Ý 00DD	Þ 00DE	ß 00DF
E0	à 00E0	á 00E1	â 00E2	ã 00E3	ä 00E4	å 00E5	æ 00E6	ç 00E7	è 00E8	é 00E9	ê 00EA	ë 00EB	ì 00EC	í 00ED	î 00EE	ï 00EF
F0	ð 00F0	ñ 00F1	ò 00F2	ó 00F3	ô 00F4	õ 00F5	ö 00F6	÷ 00F7	ø 00F8	ù 00F9	ú 00FA	û 00FB	ü 00FC	ý 00FD	þ 00FE	ÿ 00FF

È
00C8

Central Europe

Č

Arabic

ﻉ

Hebrew

ט

Greek

Θ

Cyrillic

И

Thai

ศ

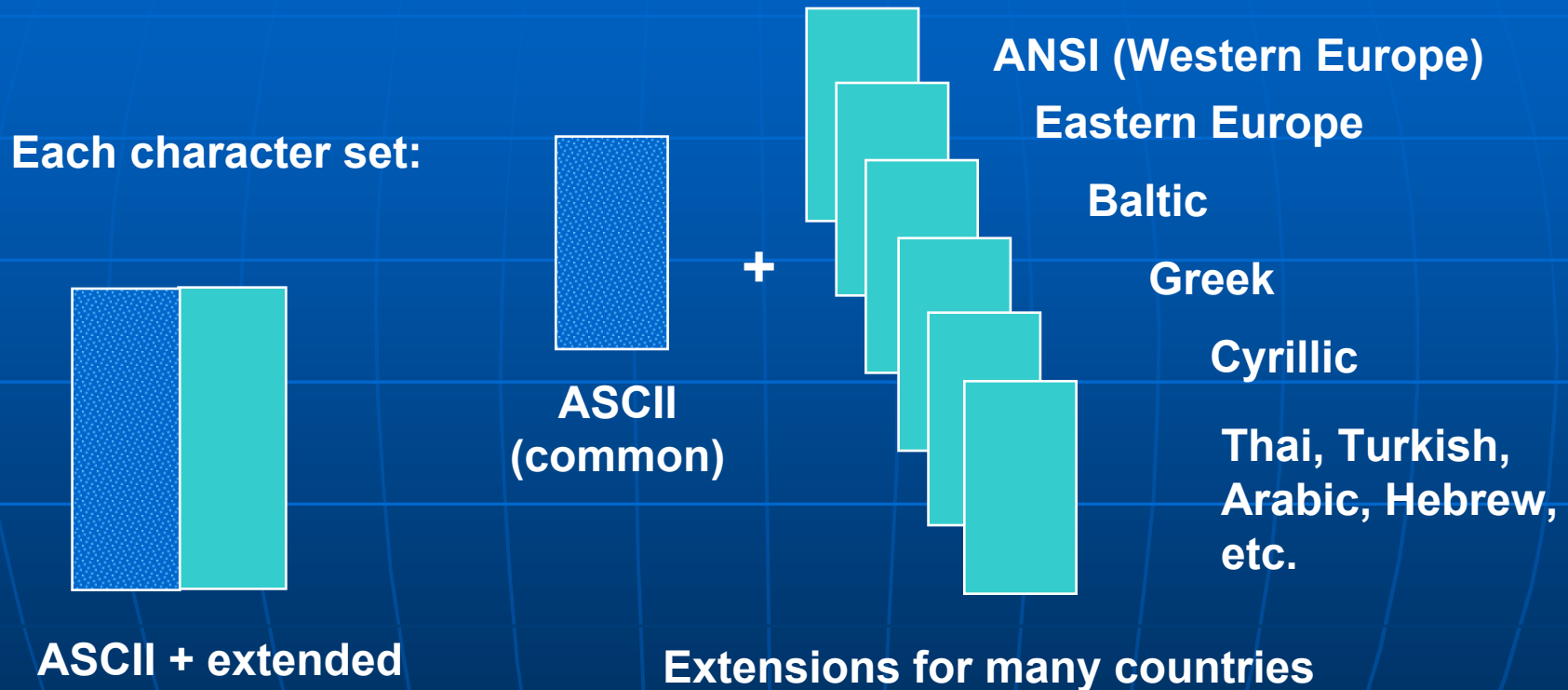
1256 WINDOWS ARABIC

	20	30	40	50	60	70	80	90	A0	B0	C0	D0	E0	F0
0	0 32	@ 48	P 64	` 80	p 96	NOT USED 112	گ 128	NBSP 144	° 160	° 176	NOT USED 192	ز 208	à 224	° 240
1	! 33	1 49	A 65	Q 81	a 97	q 113	پ 129	‘ 145	، 161	± 177	ء 193	ر 209	ل 225	° 241
2	" 34	2 50	B 66	R 82	b 98	r 114	، 130	’ 146	ø 162	² 178	آ 194	ز 210	â 226	° 242
3	# 35	3 51	C 67	S 83	c 99	s 115	f 131	“ 147	£ 163	³ 179	أ 195	س 211	م 227	° 243
4	\$ 36	4 52	D 68	T 84	d 100	t 116	„ 132	” 148	¤ 164	´ 180	ؤ 196	ش 212	ن 228	ô 244
5	% 37	5 53	E 69	U 85	e 101	u 117	... 133	• 149	¥ 165	µ 181	إ 197	ك 213	ه 229	° 245
6	& 38	6 54	F 70	V 86	f 102	v 118	† 134	- 150	! 166	¶ 182	ئ 198	ظ 214	و 230	° 246
7	' 39	7 55	G 71	W 87	g 103	w 119	‡ 135	- 151	§ 167	· 183	ا 199	x 215	ç 231	÷ 247
8	(40	8 56	H 72	X 88	h 104	x 120	^ 136	NOT USED 152	¨ 168	° 184	ب 200	ط 216	è 232	° 248
9) 41	9 57	I 73	Y 89	i 105	y 121	% 137	TM 153	© 169	¹ 185	ة 201	ظ 217	é 233	ù 249
A	* 42	: 58	J 74	Z 90	j 106	z 122	NOT USED 138	NOT USED 154	NOT USED 170	¿ 186	ت 202	ع 218	ê 234	° 250
B	+ 43	; 59	K 75	[91	k 107	{ 123	< 139	> 155	« 171	» 187	ث 203	غ 219	ë 235	û 251
C	, 44	< 60	L 76	\ 92	l 108	 124	Œ 140	œ 156	- 172	¼ 188	ج 204	ا 220	ü 236	° 252
D	- 45	= 61	M 77] 93	m 109	} 125	چ 141	Z-W N-J 157	SHY 173	½ 189	ح 205	ف 221	ي 237	L-R Z-W M-K 253
E	. 46	> 62	N 78	^ 94	n 110	~ 126	ژ 142	Z-W Jeln 158	® 174	¾ 190	خ 206	ق 222	î 238	R-L Z-W M-K 254
F	/ 47	? 63	O 79	_ 95	o 111	° 127	NOT USED 143	NOT USED 159	- 175	? 191	د 207	ك 223	ï 239	NOT USED 255

Code page
for Windows

Characters not
common to other
codepages

The Code Page Problem cont.



Characters above 128 change meaning

The Code Page Problem cont.

- Characters in most languages are traditionally represented by single-byte values
 - Allows for 256 characters max
 - Real limit for most encodings is 192 characters
 - This includes letters, digits, punctuation, symbols
- When a system is used for a new language, the encoding has to be adapted to use that language's characters

The Code Page Problem

- Each language or group of languages gets its own encoding
- Different vendors or standards committees devise different encodings, so generally each language has several, often incompatible, encodings

Interoperability Problems

- Can't easily mix languages in a document or system
- Data not tagged with encoding, so loss can occur when transferring between systems
- Most encodings are ASCII-based, so problems often not seen with English-only data
- Two possible solutions:
 - Systematic tagging of textual data with encoding ID
 - *Universal* encoding standard with all languages' characters

What is Unicode?

Unicode or Universal Code

- One Universal Code for every character
*no matter what the platform,
no matter what the program,
no matter what the language.*
- Unicode is not just a bunch of code points
- Initially it was a 2 byte code, that can support over 65,000 characters
- Unicode Standard, Version 4.0 provides codes for 96,447 characters
- Adopted by ISO as ISO 10464

Principles of the Unicode Standard

- Universality
- Efficiency
- Characters, not glyphs
- Semantics
- Plain text
- Logical order
- Unification
- Convertibility Accurate

Universality (Unicode Coverage)

- European scripts
 - Latin, Greek, Cyrillic, Armenian, Georgian, IPA
- Bidirectional (Middle Eastern) scripts
 - Hebrew, Arabic, Syriac, Thaana
- Indic (Indian and Southeast Asian) scripts
 - Devanagari, Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, Malayalam, Sinhala, Thai, Lao, Khmer, Myanmar, Tibetan, Philippine
- East Asian scripts
 - Chinese (Han) characters, Japanese (Hiragana and Katakana), Korean (Hangul), Yi
- Other modern scripts
 - Mongolian, Ethiopic, Cherokee, Canadian Aboriginal
- Historical scripts
 - Runic, Ogham, Old Italic, Gothic, Deseret
- Punctuation and symbols
 - Numerals, math symbols, scientific symbols, arrows, blocks, geometric shapes, Braille, musical notation, etc.

Characters and Glyphs

g

g

g

g

fi

fi

لا

<--

ا

ا

Plain Text

- Its is a plain text
- Its is a ***formatted*** text

Logical Data Ordering and Bidi-Algorithm

G	i	d	i	_	s	a	i	d	,	_	“	א	ם	_	א	'	ן	_	א	'	_	ל	'	_	מ	'	_	ל	'	”	.
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Gidi said, “	אם אין אני לי מי לי	”.
→	←	→

Unification

- Unicode standard avoid duplicate encoding of characters within Scripts across languages.
- In Chinese, Japanese and Korean many ideographs are common.
- The character code U+0057 “Y” is same in English, German and French

Character Semantics cont.

- The Unicode standard includes an extensive database that specifies a large number of *character properties*, including:
 - Name
 - Type (e.g., letter, digit, punctuation mark)
 - Decomposition
 - Case and case mappings (for cased letters)
 - Numeric value (for digits and numerals)
 - Combining class (for combining characters)
 - Directionality
 - Line-breaking behavior
 - Cursive joining behavior
 - For Chinese characters, mappings to various other standards and many other properties

Character semantics

- 1781;

KHMER LETTER KHA;Lo;0;L;;;;;N;;;;;

- 17BE

KHMER VOWEL SIGN OE;Mc;0;L;;;;;N;;;;;

- 17E5

KHMER DIGIT FIVE;Nd;0;L;;5;5;5;N;;;;;

Unicode Architecture

Unicode Architecture

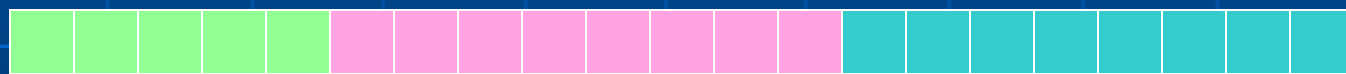
- Initially Unicode was designed for 16-bit encoding space, consisting of 256 rows of 256 characters each
- ISO 10646 was designed for 32 bit encoding space, thus ISO 10646 has room for 2,147,483,648 characters
- After Unicode came out that 16-bit encoding is too small
- In Unicode 3.0 the length is increased to 21-bit, allows for 1,114,112 characters

Encoding Space

Early versions of Unicode used 16 bits



Unicode now uses 21 bits

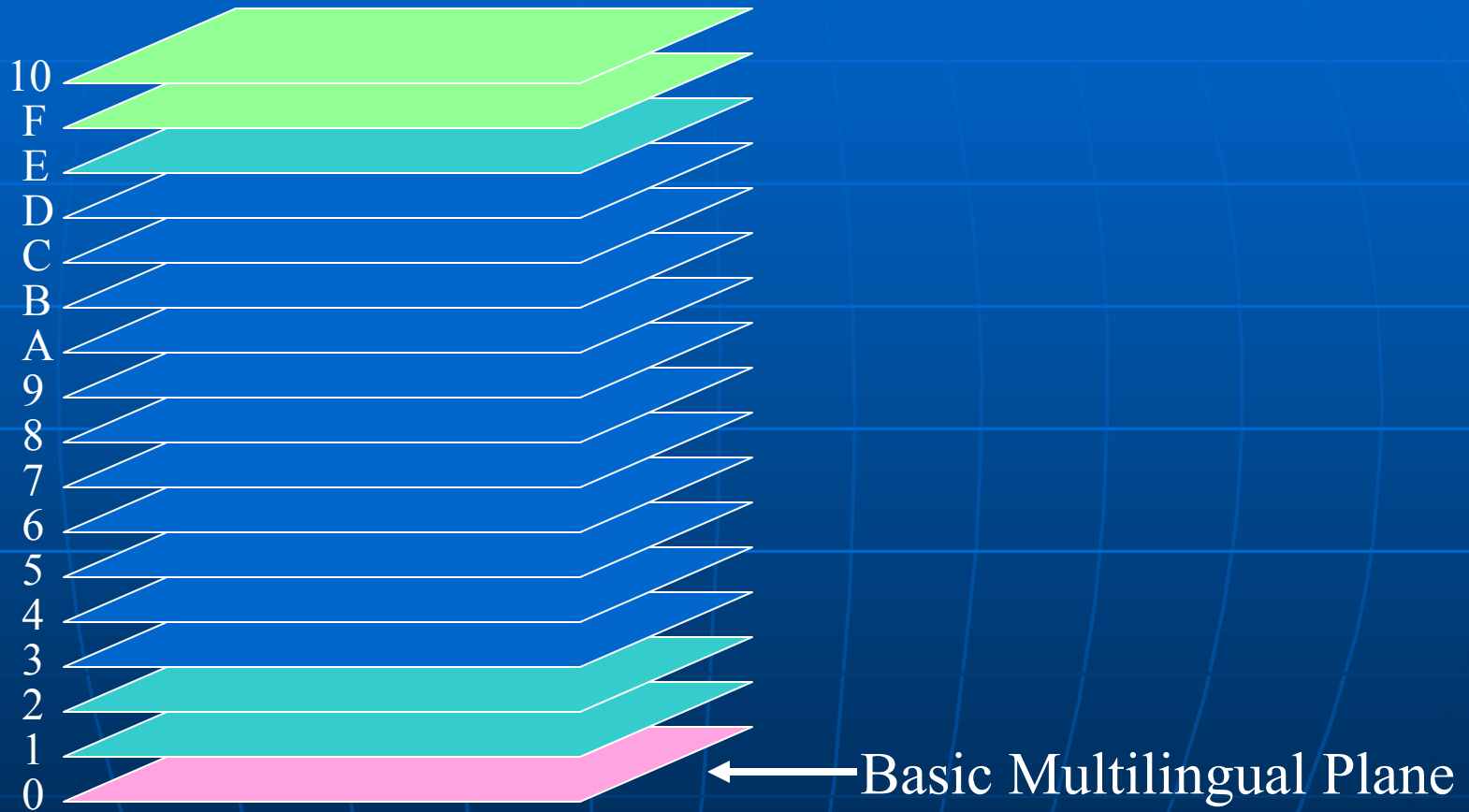


Plane
number

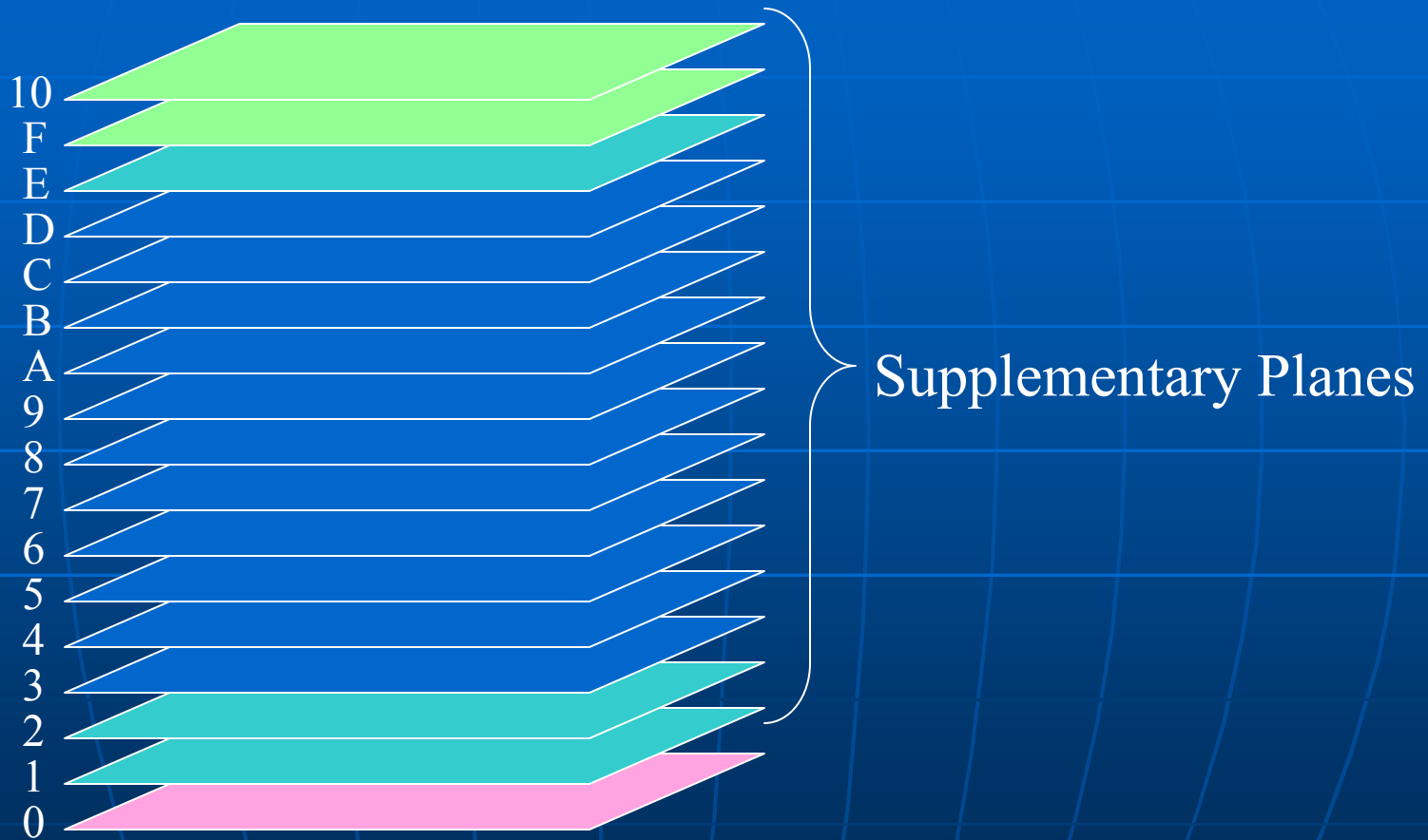
Row
number

Character
number

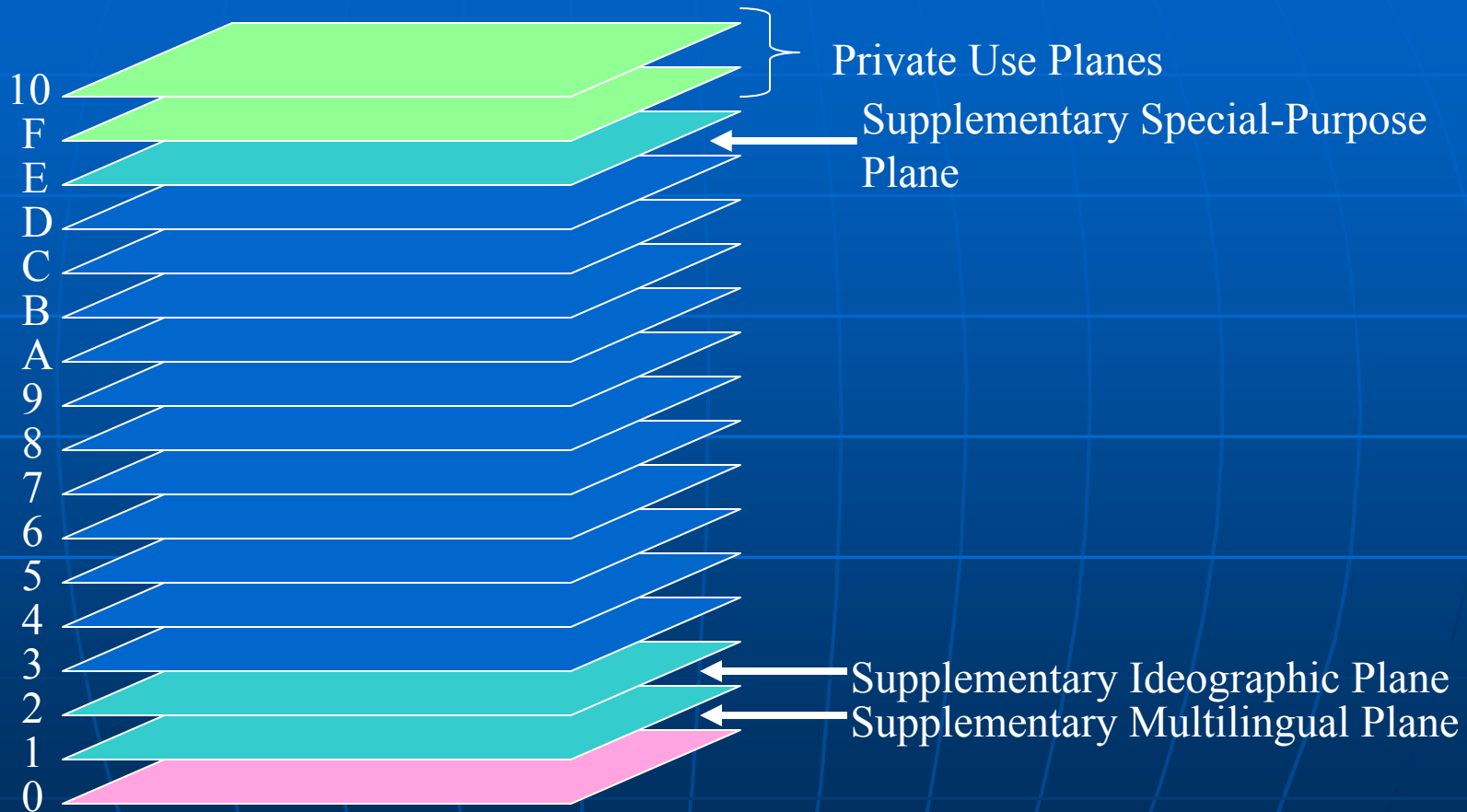
The Unicode Encoding Space



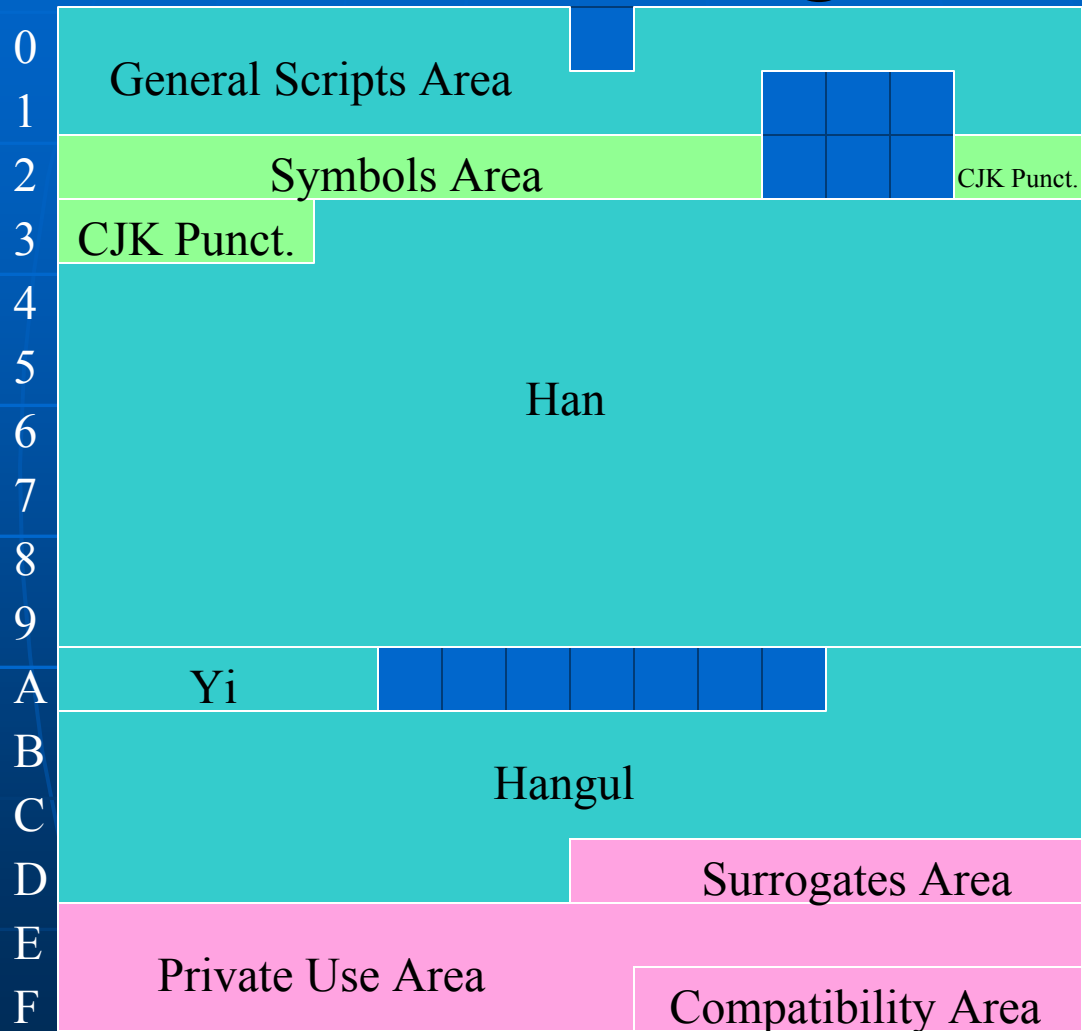
The Unicode Encoding Space



The Unicode Encoding Space



The Basic Multilingual Plane



The General Scripts Area

00/01	Latin															
02/03	IPA				Diacriticals				Greek							
04/05	Cyrillic								Armenian				Hebrew			
06/07	Arabic						Syriac				Thaana					
08/09					Devanagari				Bengali							
0A/0B	Gurmukhi				Gujarati				Oriya				Tamil			
0C/0D	Telugu				Kannada				Malayalam				Sinhala			
0E/0F	Thai				Lao				Tibetan							
10/11	Myanmar				Georgian				Hangul							
12/13	Ethiopic								Cherokee							
14/15	Canadian Aboriginal Syllabic															
16/17	Ogham		Runic				Philippine				Khmer					
18/19	Mongolian															
1A/1B																
1C/1D																
1E/1F	Latin								Greek							

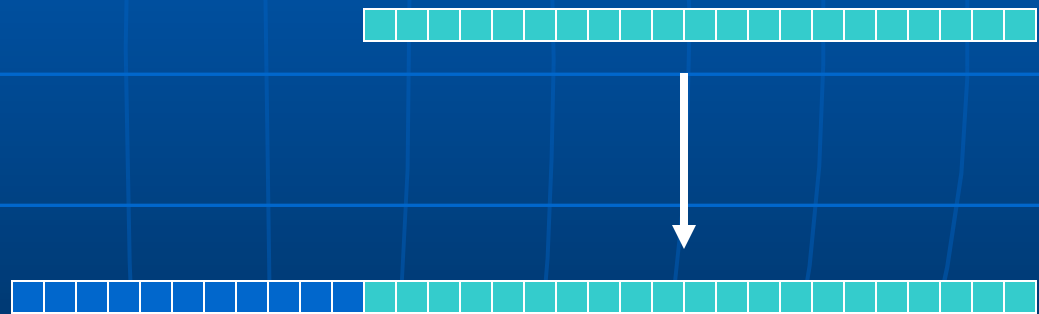
Unicode Storage Formats or Unicode Encodings

- UTF-32
- UTF-16
- UTF-8
- UTF-7
- CESU-8
- UTF-EBCDIC
- BOCU

Storage formats cont.

UTF-32:

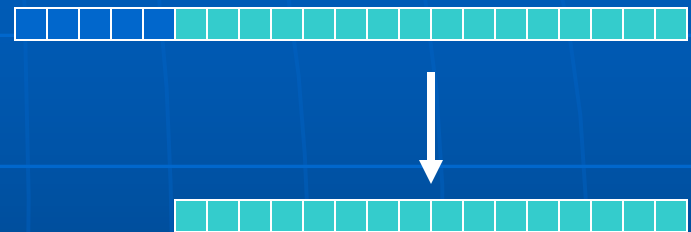
The 21-bit abstract Unicode value is simply zero-padded to 32 bits:



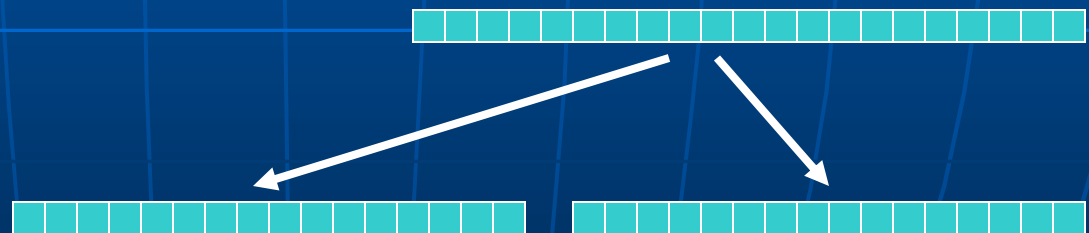
Storage formats

UTF-16:

For characters in the BMP, the 21-bit value is simply truncated to 16 bits:



For other characters, the 21-bit value is turned into a sequence of two 16-bit values called a *surrogate pair*:

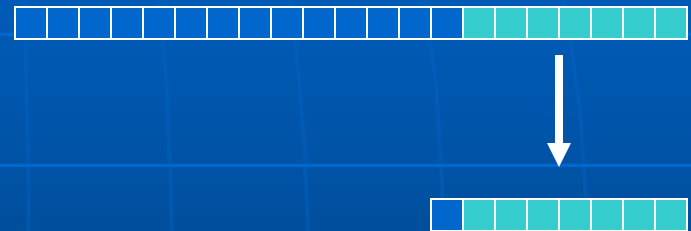


A particular numeric value is either a BMP character, a high surrogate, or a low surrogate.

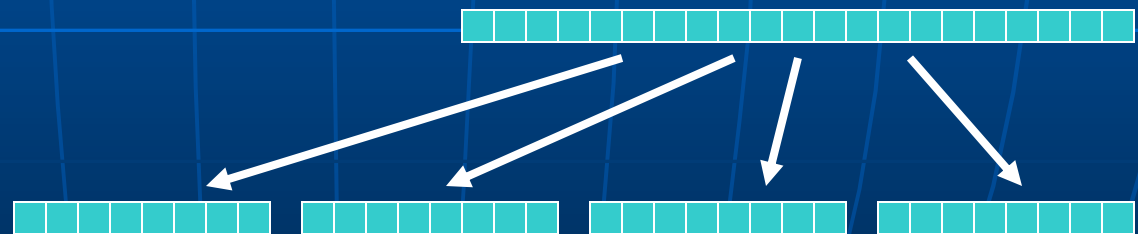
Storage formats

UTF-8:

For ASCII characters, the 21-bit value is truncated to 8 bits:



For other characters, the 21-bit value is turned into a sequence of two, three, or four 8-bit values:



Different numeric ranges are used for ASCII characters and leading and trailing bytes. Different ranges are used for leading bytes of different-length sequences.

Detecting Unicode Storage Format

If the files starts with	The file contains
0xFE 0xFF	UTF-16
0xFF 0xFE	Byte-swapped UTF-16
0x00 0x00 0xFE 0xFF	UTF-32
0xFF 0xFE 0x00 0x00	Byte-swapped UTF-32
0xEF 0xBB 0x BF	UTF-8
0xDD 0x73 0x73 0x73	UTF-EBCDIC
0x0E 0xFE 0xFE	CESU
Any thing else	Non-Unicode or Untagged Unicode

The Unicode standard

- The Unicode standard consists of:
 - The standard text, published in book form (this includes a complete set of printed code charts)
 - The Unicode Character Database, a set of data files providing complete property information on every character
 - Various Web-published supplemental materials:
 - Unicode Standard Annexes (UAX): Amendments to the standard since the last book was published
 - Unicode Technical Standards (UTS): Allied standards maintained separately from Unicode itself
 - Unicode Technical Reports (UTR): Non-normative documents providing background info, implementation hints, or other useful information
 - Unicode Technical Notes (UTN): Other articles of interest

References

- www.unicode.org
- *The Unicode Standard Version 4.0* by Unicode Consortium
- *Unicode Demystified* by Richard Gillam
- Unicode Character Database (<http://www.unicode.org/ucd/>)
- Unicode Charts (<http://www.unicode.org/charts/>)

Questions ?