WEB & INTERNATIONALIZATION

REPRESENTING WORLD WIDE WEB RESOURCES

Language	Percent of World Population
Mandarin	14.4%
Spanish	6.15%
English	5.43%
Hindi	4.70%
Arabic	4.43%
Portuguese	3.27%
Bengali	3.11%
Russian	2.33%
Japanese	1.90%
Punjabi	1.44%
German	1.39%
Javanese	1.25%

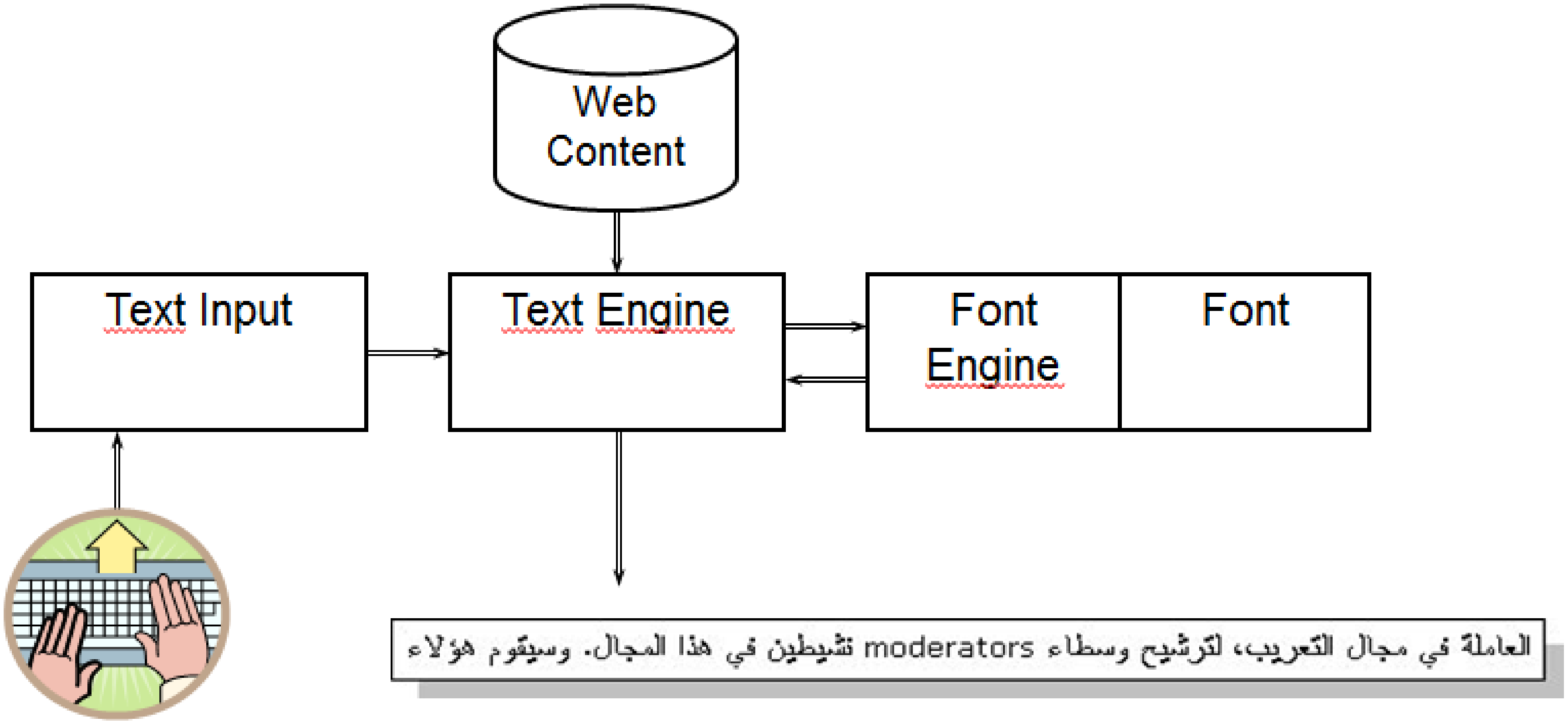
Source: Wikipedia, 11/2014

THE INTERNATIONALIZATION (I18N) PROBLEM

- Web resources are mostly text-based resources
- What is text?
  - A sequence of character
    - What is a character?
      - in English, in French, in Chinese, in Arabic …
      - what about symbols (e.g €), punctuation (., ¿) …
      - Difference character/character code (used for storage/transfer)
      - Difference character/graphical representation (used for display)
- Need for a text representation
  - Working for all languages
  - Including alphabets, ideograms, writing modes, …
  - Efficient for storage and network transfer
  - Efficient for display, editing, text selection
- Fundamentals
  - Unicode: Character Set
  - UTF-8: Encoding

I18N HANDLING

I18N PROCESSING



العاملة في مجال التعريب، لترشيح وسطاء moderators نشيطين في هذا المجال. وسيقوم هؤلاء

CHARACTER SET

- A set of ordered characters (aka Repertoire)
  - from one or more languages
  - closed (ASCII) or open (Unicode)

- Universal Character Set
  - Each character is only present once in the set
  - Characters are defined independently of their graphical representation or position in a text

- Each character is identified by its position (code position, code point)
- Characters from a set are encoded to store/transmit text: codec character set, character encoding

ASCII

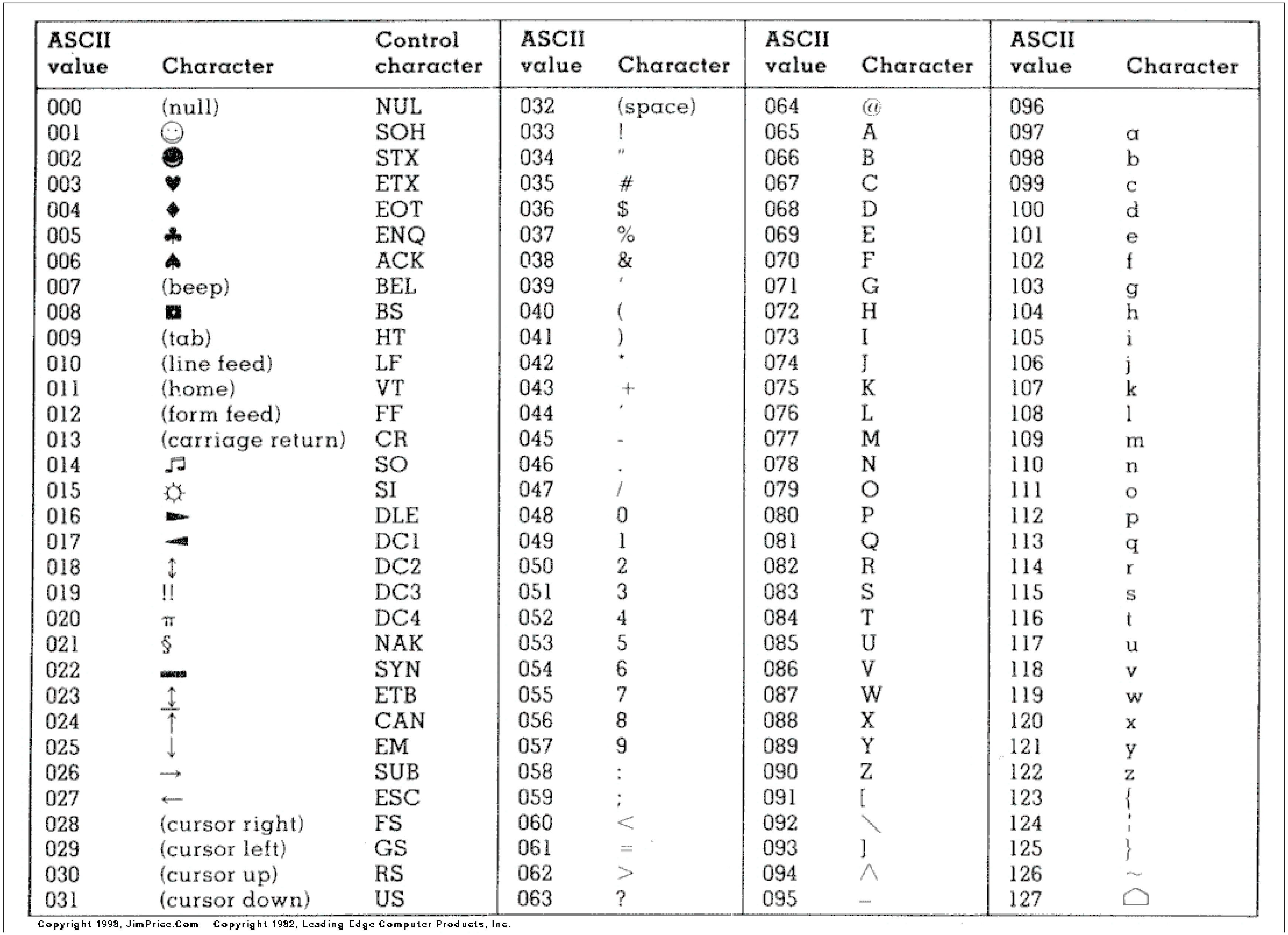
- American Standard Code for Information Interchange
  - Invented in 1965 in the USA, standardised in 1983 as ISO 646
  - Derived with many variants
  - Widely used

- Set of 128 characters
  - 33 command characters (ex CR)
  - 95 printable character
    - 83 characters common to all ASCII variants
      - small, capital roman letters
      - digits
      - punctuation: (! " % & ‘ * + , - . / : ; < = > ? _) and space

    - 2 symbols: # or £ et $ or ¤
    - 10 variable characters (per country)

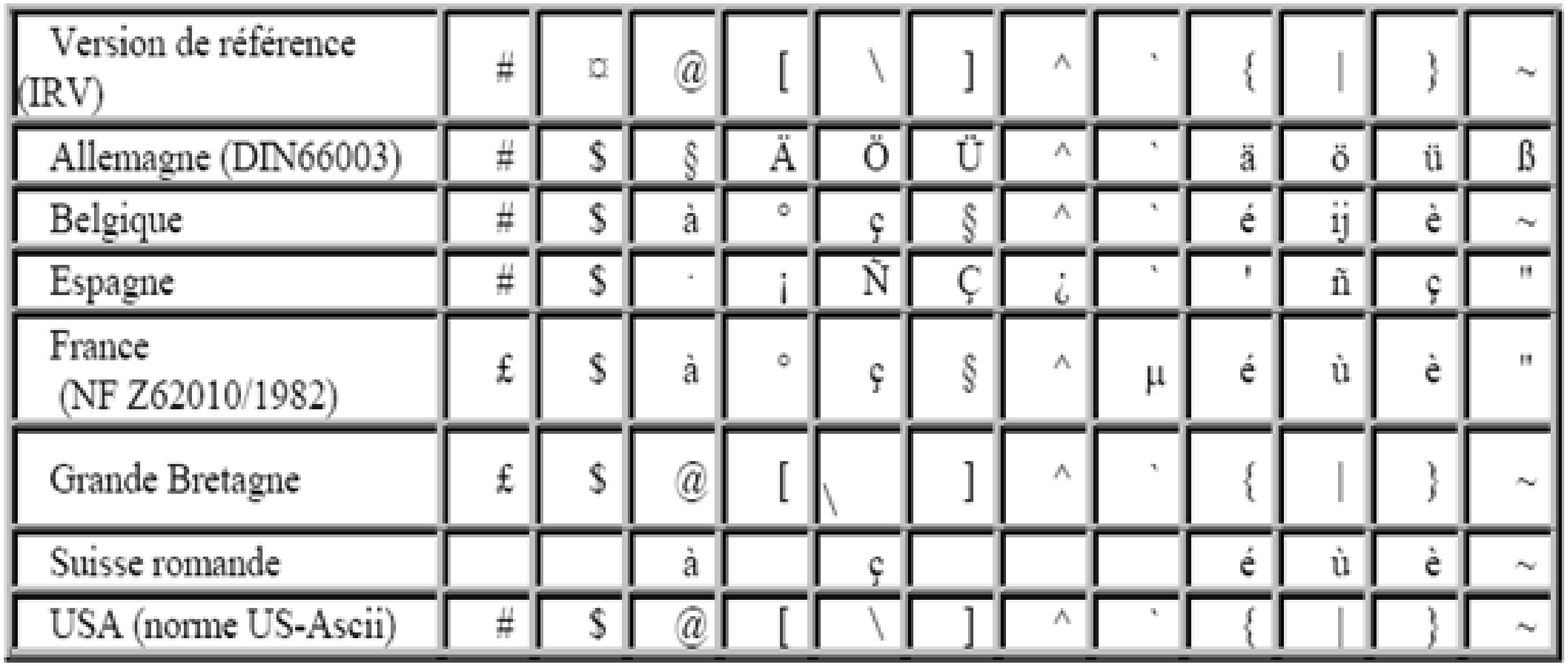
- Associated encoding on 7-bits

ASCII



ASCII value	Character	Control character	ASCII value	Character	ASCII value	Character	ASCII value	Character
000	(null)	NUL	032	(space)	064	@	096	
001	☺	SOH	033	!	065	A	097	a
002	☻	STX	034	"	066	B	098	b
003	♥	ETX	035	#	067	C	099	c
004	♦	EOT	036	$	068	D	100	d
005	♣	ENQ	037	%	069	E	101	e
006	♠	ACK	038	&	070	F	102	f
007	(beep)	BEL	039	'	071	G	103	g
008	◘	BS	040	(	072	H	104	h
009	(tab)	HT	041	)	073	I	105	i
010	(line feed)	LF	042	*	074	J	106	j
011	(home)	VT	043	+	075	K	107	k
012	(form feed)	FF	044	,	076	L	108	l
013	(carriage return)	CR	045	-	077	M	109	m
014	♫	SO	046	.	078	N	110	n
015	☼	SI	047	/	079	O	111	o
016	►	DLE	048	0	080	P	112	p
017	◄	DC1	049	1	081	Q	113	q
018	↕	DC2	050	2	082	R	114	r
019	!!	DC3	051	3	083	S	115	s
020	π	DC4	052	4	084	T	116	t
021	§	NAK	053	5	085	U	117	u
022	▬	SYN	054	6	086	V	118	v
023	↨	ETB	055	7	087	W	119	w
024	↑	CAN	056	8	088	X	120	x
025	↓	EM	057	9	089	Y	121	y
026	→	SUB	058	:	090	Z	122	z
027	←	ESC	059	;	091	[	123	{
028	(cursor right)	FS	060	<	092	\	124	¦
029	(cursor left)	GS	061	=	093	]	125	}
030	(cursor up)	RS	062	>	094	^	126	~
031	(cursor down)	US	063	?	095	_	127	⌂

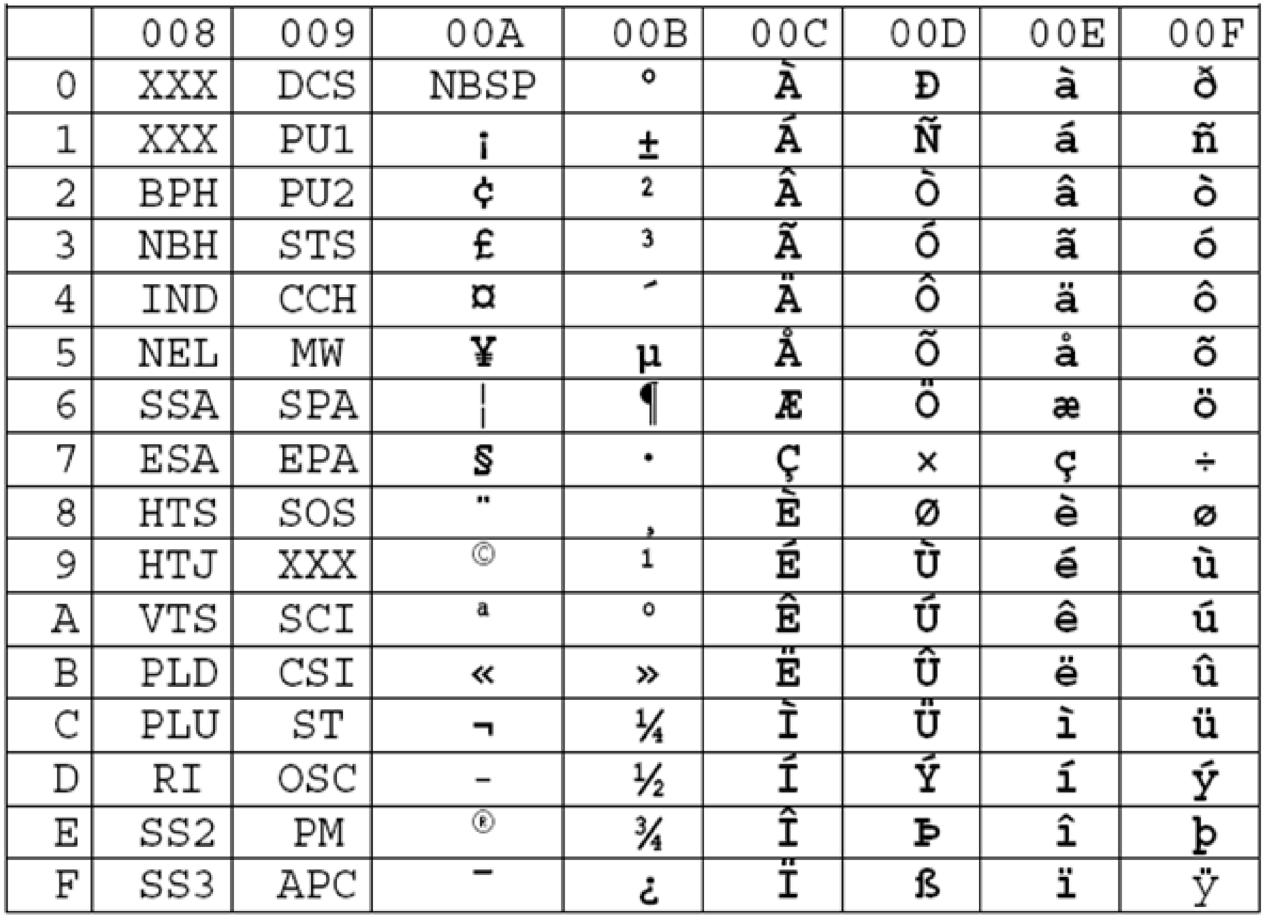
ASCII VARIANTS



Version de référence (IRV)	#	¤	@	[	\	]	^	`	{	|	}	~
Allemagne (DIN66003)	#	$	§	Ä	Ö	Ü	^	`	ä	ö	ü	ß
Belgique	#	$	à	°	ç	§	^	`	é	ij	è	~
Espagne	#	$	·	¡	Ñ	Ç	¿	`	'	ñ	ç	"
France (NF Z62010/1982)	£	$	à	°	ç	§	^	µ	é	ù	è	"
Grande Bretagne	£	$	@	[	\	]	^	`	{	|	}	~
Suisse romande			à		ç				é	ù	è	~
USA (norme US-Ascii)	#	$	@	[	\	]	^	`	{	|	}	~

ISO-8859

- 8-bit extension to ASCII
- Same 128 first characters as ASCII
- 32 additional characters
- 96 language-specific characters
- ISO/IEC 8859-n, n=1…16 (aka Latin-1, Latin-2 …)

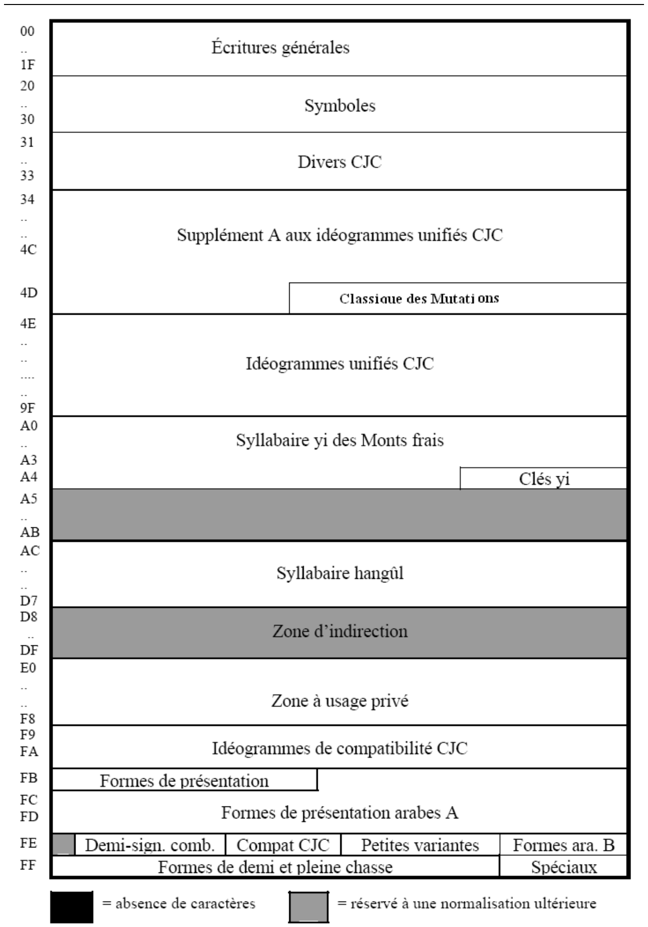


	008	009	00A	00B	00C	00D	00E	00F
0	XXX	DCS	NBSP	°	À	Ð	à	ð
1	XXX	PU1	¡	±	Á	Ñ	á	ñ
2	BPH	PU2	¢	²	Â	Ò	â	ò
3	NBH	STS	£	³	Ã	Ó	ã	ó
4	IND	CCH	¤	´	Ä	Ô	ä	ô
5	NEL	MW	¥	µ	Å	Õ	å	õ
6	SSA	SPA	¦	¶	Æ	Ö	æ	ö
7	ESA	EPA	§	·	Ç	×	ç	÷
8	HTS	SOS	¨	¸	È	Ø	è	ø
9	HTJ	XXX	©	¹	É	Ù	é	ù
A	VTS	SCI	ª	º	Ê	Ú	ê	ú
B	PLD	CSI	«	»	Ë	Û	ë	û
C	PLU	ST	¬	¼	Ì	Ü	ì	ü
D	RI	OSC	-	½	Í	Ý	í	ý
E	SS2	PM	®	¾	Î	Þ	î	þ
F	SS3	APC	¯	¿	Ï	ß	ï	ÿ

THE UNICODE STANDARD

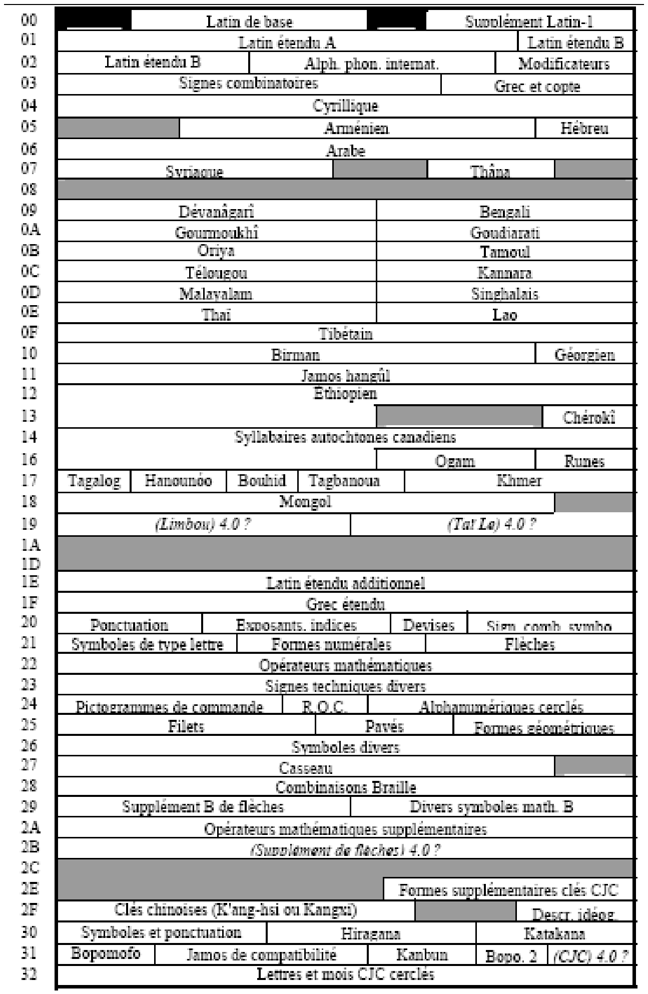
- Universal Character Set
  - More than 1 million of representable characters
- Latest version
  - Unicode 8.0 - 06/2015
  - Over 120 000 characters defined
- Grouped in 17 planes de 2^16 characters
  - Base Multilingual Plane (BMP)
  - Supplementary Multilingual Plane (SMP)
  - …

BASIC MULTILINGUAL PLANE



00 .. 1F	Écritures générales				
20 .. 30	Symboles				
31 .. 33	Divers CJC				
34 .. .. 4C	Supplément A aux idéogrammes unifiés CJC				
4D		Classique des Mutati ons			
4E 9F	Idéogrammes unifiés CJC				
A0 .. A3	Syllabaire yi des Monts frais				
A4		Clés yi			
A5 .. AB	▨				
AC D7	Syllabaire hangûl				
D8 .. DF	Zone d'indirection				
E0 .. .. F8	Zone à usage privé				
F9 FA	Idéogrammes de compatibilité CJC				
FB	Formes de présentation				
FC FD	Formes de présentation arabes A				
FE	▨	Demi-sign. comb.	Compat CJC	Petites variantes	Formes ara. B
FF	Formes de demi et pleine chasse	Spéciaux			

■ = absence de caractères ▨ = réservé à une normalisation ultérieure



00	■	Latin de base	■	Supplément Latin-1	
01	Latin étendu A	Latin étendu B			
02	Latin étendu B	Alph. phon. internat.	Modificateurs		
03	Signes combinatoires	Grec et copte			
04	Cyrillique				
05	▨	Arménien	Hébreu		
06	Arabe				
07	Syriaque	▨	Thâna	▨	
08	▨				
09	Dévanâgarî	Bengali			
0A	Gourmoukhî	Goudjarati			
0B	Oriya	Tamoul			
0C	Télougou	Kannara			
0D	Malayalam	Singhalais			
0E	Thai	Lao			
0F	Tibétain				
10	Birman	Géorgien			
11	Jamos hangûl				
12	Éthiopien				
13		▨	Chérokî		
14	Syllabaires autochtones canadiens				
16		Ogam	Runes		
17	Tagalog	Hanounóo	Bouhid	Tagbanoua	Khmer
18	Mongol	▨			
19	*(Limbou) 4.0 ?*	*(Taï Le) 4.0 ?*			
1A 1D	▨				
1E	Latin étendu additionnel				
1F	Grec étendu				
20	Ponctuation	Exposants, indices	Devises	Sign. comb. symbo	
21	Symboles de type lettre	Formes numérales	Flèches		
22	Opérateurs mathématiques				
23	Signes techniques divers				
24	Pictogrammes de commande	R.O.C.	Alphanumériques cerclés		
25	Filets	Pavés	Formes géométriques		
26	Symboles divers				
27	Casseau	▨			
28	Combinaisons Braille				
29	Supplément B de flèches	Divers symboles math. B			
2A	Opérateurs mathématiques supplémentaires				
2B	*(Supplément de flèches) 4.0 ?*				
2C	▨				
2E	▨	Formes supplémentaires clés CJC			
2F	Clés chinoises (K'ang-hsi ou Kangxi)	▨	Descr. idéog.		
30	Symboles et ponctuation	Hiragana	Katakana		
31	Bopomofo	Jamos de compatibilité	Kanbun	Bopo. 2	*(CJC) 4.0 ?*
32	Lettres et mois CJC cerclés				

A UNICODE CODE POINT

- Each character is assigned
  - A unique code point (code position):
    - U+xxxx (BMP) Ex: U+0044
    - Ex : U+yyxxxx (other planes)
  - A name: ex Capital latin letter D
  - A direction: « left – right » or « right – left »
  - A possible decomposition : é=e + ‘
  - Some language information
- The graphical shape is not associated
  - see Font information
- The byte representation on the wire is not defined in Unicode
  - see Character Encoding (fixed length, variable length)

FIXED-LENGTH CHARACTER ENCODING

- Mostly defined by ISO
- ASCII
  - Not capable of encoding the Unicode Character Set

- UCS-2 (deprecated)
  - 16 bits - PMB
  - Not ASCII-compatible

- UCS-4 (deprecated)
  - 31 bits (+ leading 0 bit)
  - Designed for 32-bits machines
  - Restricted to [0x0..0x10FFFF] for UTF-16 compatibility
  - Not ASCII-compatible

VARIABLE LENGTH CHARACTER ENCODINGS

- Mostly defined by IETF (RFC 2279, 1998)
- UTF-8: Universal Transformation Format
  - Most popular format
  - 1-Byte alignment (no multi-byte problem)
  - ASCII-compatible (0..127)
    - An ASCII file transcoded in UTF-8 is identical to the original file
    - Bytes with the most-significant bit set to 1 are ignored by ASCII processors
  - Efficient conversion into UTF-16 & UTF-32
  - Used in Java
- UTF-16
  - Alignment on 2-bytes
  - BMP=2 bytes
  - Other planes=2 (indirection) + 2
  - Use of Byte Order Mark (BOM) to detect Endianness
  - Used on Windows
- UTF-32=UCS-4

UNIVERSAL TRANSFORMATION FORMAT

Code Position Unicode	UTF-16	UTF-8 1st byte	UTF-8 2nd byte	UTF-8 3rd byte	UTF-8 4th byte
0000 0000 0xxx xxxx	0000 0000 0xxx xxxx	0xxx xxxx			
0000 0yyy yyxx xxxx	0000 0yyy yyxx xxxx	110y yyyy	10xx xxxx		
zzzz yyyy yyxx xxxx	zzzz yyyy yyxx xxxx	1110 zzzz	10yy yyyy	10xx xxxx	
000u uuuu zzzz yyyy yyxx xxxx	1101 10ww wwzz zzyy + 1101 11yy yyxx xxxx wwww=uuuuu–1	1111 0uuu	10uu zzzz	10yy yyyy	10xx xxxx

UNICODE & ENCODINGS

EXAMPLE AND COUNTER-EXAMPLES

Character	Unicode Code	UTF-8	UTF-8 in ASCII	UTF-16 (BE)	UTF-16 (LE)	UTF-32
A	U+0041	41	A	0041	4100	0000 0041
space	U+0020	20		0020	2000	0000 0020
é	U+00C9	C3 A9	Ã©	00E9	E900	0000 00E9
δ	U+03B4	CE B4	Î´	03B4	B403	0000 03B4
Å	U+00C5	C3 85	Ã…	00C5	C500	0000 00C5
Å	U+212B	E2 84 AB	â„«	212B	2B21	0000 212B
A + °	U+0041 + U+030A	41 CC 8A	AÌŠ	0041 030A	4100 0A03	0000 0041 0000 030A

OTHER ENCODINGS

- ISO-8859-1: Western Europe
- ISO-8859-6: Arabic
- ISO-8859-11: Thai
- Windows-1252: Western languages
- Shift-JIS: Japanese
- GB-2312: Chinese Guobiao
- Big-5: Taïwan
- ISO-2022-KR: Korean
- …

DECLARING CHARACTER ENCODING

- In HTTP Headers

```
Content-Type: ISO-8859-1
```

- XML Declaration

```
<?xml version="1.0" encoding="ISO-8859-1"?>
```

- In HTML Documents

```
<meta charset='utf-8'>
<meta http-equiv="Content-Type" content="text/html;charset=UTF-8" />
```

ESCAPE CODES IN WEB CONTENT

Character(s)	é	Å	δ	±	space	Text
HTML Escaping	&acute; / &#x00C9;	&Aring; / &#x212B;	&delta; / &#x03B4;	&plusmn; / &#x00B1;	&nbsp; / &#x0020;	Text
URL escaping	%C3%A9	%C3%85	%CE%B4	%C2%B1	%20	Text
Base 64 encoding	w6k=	w4U=	zrQ=	wrE=	IA==	VGV4dA==
MIME Escaping	=C3=A9	=C3=85	=CE=B4	=C2=B1	=	Text

Online encoder/decoder

Next to fonts