План занятия (20.10.17)

- Unicode

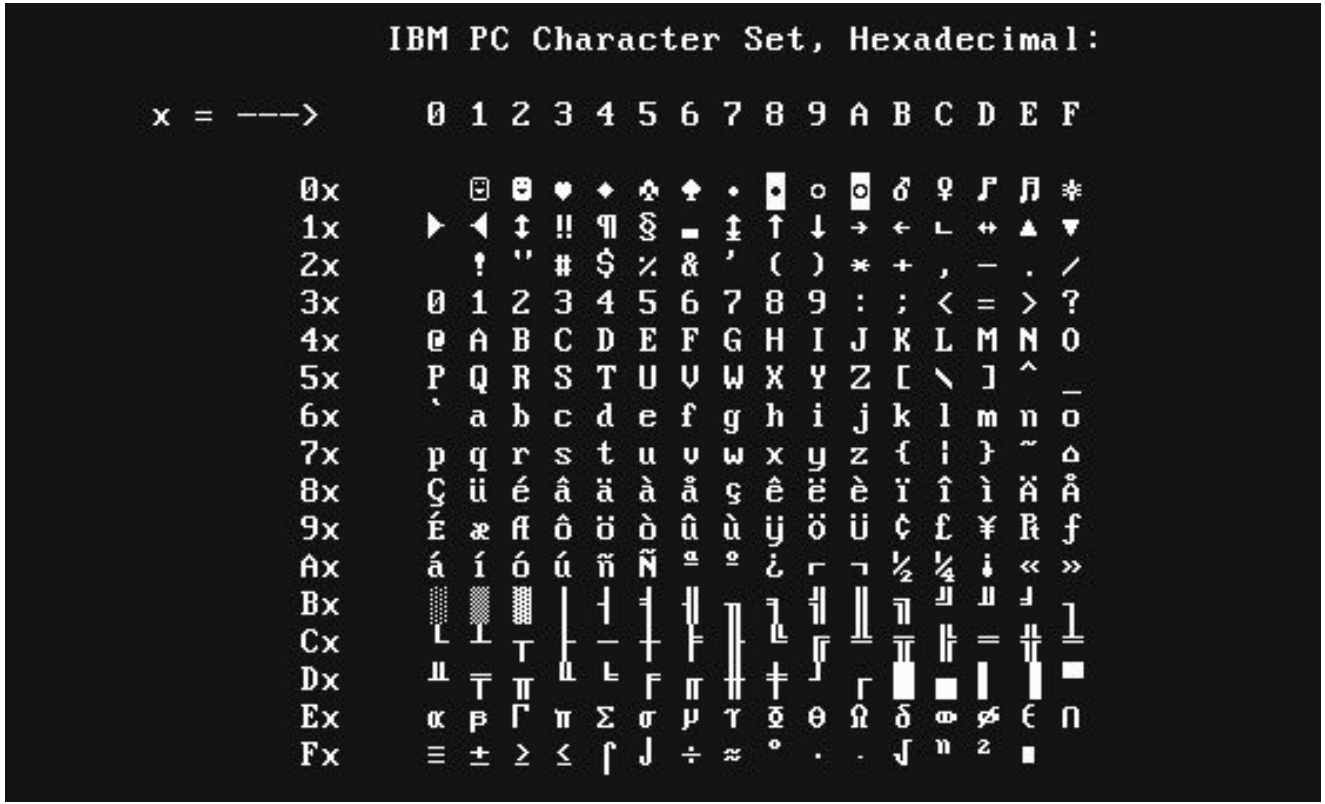
ASCII (1963)

ASCII = American standard code for information interchange

ASCII Code Chart

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

OEM Character Set (by IBM)



Проблема

130 = é

Résumés

130 = ג

Rג sumג s

Решение ?

Code pages:

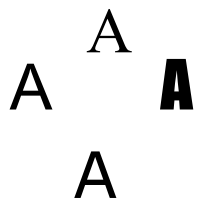
http://www.i18nguy.com/unicode/codepages.html

Но что делать, если:

- Хотим писать сразу на нескольких языках
- CJK - Chinese, Japanese, Korean
  - решение - Double Byte Character Set

Unicode

Letter



Code Point

U+0041

Unicode ничего не говори о том, как символ будет представляться в памяти!

https://unicode-table.com/ru/

Encodings (UCS-2)

Hello U+0048 U+0065 U+006C U+006C U+006F

Как можно закодировать?

- 00 48 00 65 00 6C 00 6C 00 6F
- 48 00 65 00 6C 00 6C 00 6F 00

Для задания порядка - Byte Order Mark:

- FE FF
- FF FE

Encodings (UCS-2)

Все хорошо? - Не совсем:

- неэффективно по памяти
- кто будет конвертировать старые файлы в ASCII

Encodings (UTF-8)

Число байт	Бит на code point	Первая code point	Последняя code point	Байт 1	Байт 2	Байт 3	Байт 4
1	7	U+0000	U+007F	0xxxxxxx			
2	11	U+0080	U+07FF	110xxxxx	10xxxxxx		
3	16	U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx	
4	21	U+10000	U+10FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

Encodings

- UTF-16: одно или два 16-ти битных слова
- UTF-32: 32 бита