Character Representation

Representing Text

- To represent a text document in digital form, we need to be able to represent every possible character that may appear.
- There are finite number of characters to represent, so the general approach is to list them all and assign each a binary string.
- A **character set** is a list of characters and the codes used to represent each one.
- By agreeing to use a **particular character** set, computer manufacturers have made the processing of text data easier.

Character Storage Systems

- Character sets
  - Standard ASCII (0 – 127)
  - Extended ASCII (0 – 255)
  - ANSI (0 – 255)
  - Unicode (0 – 65,535)
- Null-terminated String
  - Array of characters followed by a *null byte*

The ASCII Character Set

- ASCII stands for American Standard Code for Information Interchange. The ASCII character set originally used seven bits to represent each character, allowing for 128 unique characters.

The ASCII Character Set

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

The ASCII character set

	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	SPC	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
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

- CR = “carriage return” (MSDOS: move to beginning of line)
- LF = “line feed” (MSDOS: move directly one line below)

- SPC = “blank space”

The ASCII Character Set

- Note that the first 32 characters in the ASCII character chart do not have a simple character representation that you could print to the screen.

ASCII

- 0 – 31 and 127= unprintable
- 32 – 126 = Printable

- Computers could use 8 bits, ASCII only used 7 bits.
- Some people thought:
- “We can use 128-255 for whatever we want!”.
  - Parity Checking
  - IBM-PC
    - OEM Character Set provided accented characters for European Languages
  - More and more users were using the top 128 characters for their own purposes
  - Example:
    - On some PCs the character code 130 would display **é**
    - Computers sold in Israel it was the Hebrew letter **ג**
    - So when Americans sending their **résumés** to Israel they would arrive as **rגsumגs**

ASCII vs Extended ASCII

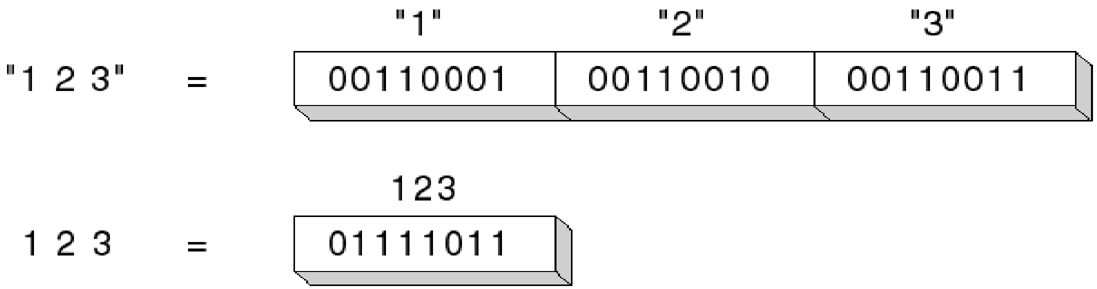
- The ASCII code (from 00h to 7Fh)
  - Only codes from 20h to 7Eh represent printable characters. The rest are control codes (used for printing, transmission…).
- Extended ASCII character set (codes 80h to FFh)
  - Varies from one system to another
    - MS-DOS usage: for accentuated characters, Greek symbols and some graphic characters

Text Files

- These are files containing only ASCII characters
- But different conventions are used for indicating an “end-of line”
  - MS-DOS: <CR>+<LF>
  - UNIX: <LF>
  - MAC: <CR>
- This is at the origin of many problems encountered during transfers of text files from one system to another

Strings and numbers

- A strings is stored as an array of characters
- A 1-byte ASCII code is stored for each char
- Hence, we can either store the number 123 in numerical form or as the string “123”
  - The string form is best for display
  - The numerical form is best for computations



"1" "2" "3"

"1 2 3" = 00110001 00110010 00110011

123

1 2 3 = 01111011

The Unicode Character Set

- The extended version of the ASCII character set is not enough for international use.
- The Unicode character set uses 16 bits per character. Therefore, the Unicode character set can represent 216, or over 65 thousand, characters.
- Unicode was designed to be a superset of ASCII. That is, the first 256 characters in the Unicode character set correspond exactly to the extended ASCII character set.

The Unicode Character Set

Code (Hex)	Character	Source
0041	A	English (Latin)
042F	Я	Russian (Cyrillic)
OE09	ฉ	Thai
13EA	Ꮺ	Cherokee
211E	℞	Letterlike Symbols
21CC	⇌	Arrows
282F	⠯	Braille
345F	㑟	Chinese/Japanese/ Korean (Common)

Figure 3.6 A few characters in the Unicode character set