ICT Foundation

Character representation

Copyright © 2010, IT Gatekeeper Project – Ohiwa Lab. All rights reserved.

Using binary numbers to represent characters

- Computer can handle character data
  - For example, mapping each binary number to each character and then handling the binary numbers instead of characters.

Character	A	B	C	D
Binary number	00	01	10	11

※ These mapping rules are not the real ones in computer.

  - The process of converting characters to binary strings (bit sequences) which computer can handle is called encoding.
  - The table that consists characters and binary strings is called character code table.

Copyright © 2010, IT Gatekeeper Project – Ohiwa Lab. All rights reserved.

Character Encoding
ASCII code

- ASCII code
  - Characters are encoded in 7-digit binary numbers.
    - The 4 left-most bits represent hexadecimal numbers from 0~F, the 3 right-most bits represent hexadecimal numbers from 0~7. (E.g.: $A = 41_{(16)} = 1000001_{(2)}$)
  - Specially defined control characters such as CR, DEL, etc. do specific functions.
    - BS (Back Space): turn back by one character
    - CR (Carriage Return): start new line of text
  - Japanese uses many characters so that 7 bits can not represent sufficiently.

Copyright © 2010, IT Gatekeeper Project – Ohiwa Lab. All rights reserved.

ASCII code table

	0	1	2	3	4	5	6	7
0	Null	DLE	Space	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

Copyright © 2010, IT Gatekeeper Project – Ohiwa Lab. All rights reserved.

Japanese Encoding
Multibyte Encoding (variable-width encoding)

- Japanese including Kanji with 65536 characters can be represented by 16-bit (2 bytes) binary numbers.
  - JIS X 0208 standard prescribes 6879 characters of Hiragana, Katakana, Kanji, …

- There are three types of encoding based on JIS X 0208
  - ISO-2022-JP（JIS）・・・ Mainly used in email
  - Shift_JIS・・・ First used in Windows and widely used in personal computer.
  - EUC-JP・・・ Mainly used in Unix

Copyright © 2010, IT Gatekeeper Project – Ohiwa Lab. All rights reserved.

Unicode

- Representation and handling of text expressed in most of the world's writing system.
  - Shift-JIS or EUC-JP that is based on JIS X 0208 is only used in Japan
  - Started to express the characters in the whole world by 16-bit binary number.
- Two encoding types
  - UCS-2, UCS-4
  - UTF-7, UTF-8, UTF-16, UTF-32
- Official website : http://unicode.org

There's also a problem of integrated working (be assumed as the same) in such languages using similar Kanji as Japanese, Chinese or Korean

Copyright © 2010, IT Gatekeeper Project – Ohiwa Lab. All rights reserved.