ICT Foundation

文字の表現

文字を2進数で表現する

- コンピュータは文字データを扱うことができる
  - 例えば，以下のように文字を2進数に対応付け（マッピング）すれば，文字を2進数として扱うことができる

文字	A	B	C	D
2進数	00	01	10	11

※ 実際にコンピュータで使われているマッピングとは異なります

  - 文字などの情報をコンピュータで処理できる符号（2進数）で表現しなおすことをコード化（符号化）という
  - 文字と符号の対応を表現した表を文字コード表という

アルファベットの符号化
ASCIIコード

- ASCIIコード
  - 文字情報を7桁の2進数に符号化
    - 上側の0～7の16進数は上位3桁，左側の0～Fの16進数は下位4桁を表現（例：A＝$41_{(16)}$＝$1000001_{(2)}$）
  - CRやDEL等はコンピュータに特定の機能を実行させる役割が割り当てられた機能コード
    - BS（Back Space）は「一文字後退」
    - CR（Carriage Return）は「行頭復帰（カーソルを，カーソルが位置する行の先頭に移動させる）」
  - 日本語はもっと多くの文字があるので，7桁では表現できない

ASCIIコード表

	0	1	2	3	4	5	6	7
0	Null	DLE	空白	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

日本語の符号化（マルチバイトコード系）

- 漢字を含む日本語は65536通りを区別できる16桁以上の2進数で表現する
  - JIS X 0208という規格で，計6879個の文字（ひらがな，カタカナ，漢字など）が規定されている
- よく使われているJIS X 0208に基づいた符号化方式は以下の3種類
  - ISO-2022-JP（JIS）・・・電子メールで主に使われている
  - Shift_JIS・・・Windowsをはじめ多くのパソコンで使われている
  - EUC-JP・・・Unixで主に使われている

Unicode

- 世界中の主要な言語の多様な文字を1つの文字コード体系で取り扱い，多言語表記を可能にする
  - JIS X 0208に基づくShift-JISやEUC-JPは日本専用
  - 全世界の文字を2進数16桁で表現しようと開始された
- 符号化方式としては以下のようなものがある
  - UCS-2，UCS-4
  - UTF-7，UTF-8，UTF-16，UTF-32
- 公式ホームページ：http://unicode.org

中国語や日本語，韓国語で使われる漢字で字形が似ている文字を同一とみなす（統合作業）などの問題点もある