文字コード

はしもとじょーじ

計算機で文字を扱う

文字コード (character code)

コンピュータでは文字や記号ひとつひとつに
固有の符号を割り当てている
文字コードとはその対応づけのこと

コンピュータが扱うのは０と１だけ
ひとかたまりの０と１の並びを文字に変換
する際の変換表が文字コード

文字コード

ASCII コード（アスキーコード）

- American Standard Code for Information Interchange
- 7 bit の文字コード(最大128文字)
- 英数字だけを使うならこれで間に合う

例：	0100011	#	0100100	$
	1000001	A	1100001	a

ASCII コード

文字	コード		文字	コード		文字	コード		文字	コード		文字	コード		文字	コード		文字	コード	
	10進	16進		10進	16進		10進	16進		10進	16進		10進	16進		10進	16進		10進	16進
NUL	0	0x00	DLE	16	0x10	SP	32	0x20	0	48	0x30	@	64	0x40	P	80	0x50	`	96	0x60	p	112	0x70
SOH	1	0x01	DC1	17	0x11	!	33	0x21	1	49	0x31	A	65	0x41	Q	81	0x51	a	97	0x61	q	113	0x71
STX	2	0x02	DC2	18	0x12	"	34	0x22	2	50	0x32	B	66	0x42	R	82	0x52	b	98	0x62	r	114	0x72
ETX	3	0x03	DC3	19	0x13	#	35	0x23	3	51	0x33	C	67	0x43	S	83	0x53	c	99	0x63	s	115	0x73
EOT	4	0x04	DC4	20	0x14	$	36	0x24	4	52	0x34	D	68	0x44	T	84	0x54	d	100	0x64	t	116	0x74
ENQ	5	0x05	NAK	21	0x15	%	37	0x25	5	53	0x35	E	69	0x45	U	85	0x55	e	101	0x65	u	117	0x75
ACK	6	0x06	SYN	22	0x16	&	38	0x26	6	54	0x36	F	70	0x46	V	86	0x56	f	102	0x66	v	118	0x76
BEL	7	0x07	ETB	23	0x17	'	39	0x27	7	55	0x37	G	71	0x47	W	87	0x57	g	103	0x67	w	119	0x77
BS	8	0x08	CAN	24	0x18	(	40	0x28	8	56	0x38	H	72	0x48	X	88	0x58	h	104	0x68	x	120	0x78
HT	9	0x09	EM	25	0x19	)	41	0x29	9	57	0x39	I	73	0x49	Y	89	0x59	i	105	0x69	y	121	0x79
NL*	10	0x0a	SUB	26	0x1a	*	42	0x2a	:	58	0x3a	J	74	0x4a	Z	90	0x5a	j	106	0x6a	z	122	0x7a
VT	11	0x0b	ESC	27	0x1b	+	43	0x2b	;	59	0x3b	K	75	0x4b	[	91	0x5b	k	107	0x6b	{	123	0x7b
NP	12	0x0c	FS	28	0x1c	,	44	0x2c	<	60	0x3c	L	76	0x4c	\	92	0x5c	l	108	0x6c	|	124	0x7c
CR	13	0x0d	GS	29	0x1d	-	45	0x2d	=	61	0x3d	M	77	0x4d	]	93	0x5d	m	109	0x6d	}	125	0x7d
SO	14	0x0e	RS	30	0x1e	.	46	0x2e	>	62	0x3e	N	78	0x4e	^	94	0x5e	n	110	0x6e	~	126	0x7e
SI	15	0x0f	US	31	0x1f	/	47	0x2f	?	63	0x3f	O	79	0x4f	_	95	0x5f	o	111	0x6f	DEL	127	0x7f

IT 用語辞典 e−Words　http://e−words.jp/p/r−ascii.html

日本語の文字コード

日本語は文字集合が大きい

- 2 byte のコード(最大65536文字)

参考：

塚本真也，日本人学生のための日本語教育
日本で使われている漢字は約５万字
常用漢字は約2000字

日本語の文字コード

面倒なことに複数の文字コードが使われている
(代表的なものだけでも 3 つある)

junet (iso-2022-jp)

- JIS コードと呼ばれたりもする

EUC-JP

- UNIX での標準 (Extended Unix Code)
  EUC-KR, EUC-CN, EUC-TW などもある

Shift_JIS

- Windows, Mac (OS9まで) での標準

文字化け

文字コードの指定の間違い

フォントの問題(機種依存文字など)

プログラムの問題

データの欠落や変質

Unicode

世界中の文字をひとつの文字コードで表す

- 当初は 2 byte で文字集合を定義
- Unicode 2.0 以降では 4 byte

- Windows NT 系、Mac OS X での標準
- UNIX 系でも使えるようになってきた

- いろいろな問題

  使う文字を決めるのは誰なのか

文字コードの相互変換

文字コードの変換ツール

- UNIX　　nkf, iconv, ...
- Windows　QKC, KanjiTranslator, ...

多言語を扱えるエディタ

- UNIX　　VIM, Emacs, ...
- Windows　VIM, Meadow, ...

改行コード

「行の終端」を表す文字(制御文字)

- CR (キャリッジリターン)
- LF (ラインフィード)

OS によって改行を表す方式が異なっている

- UNIX 系　　LF のみ
  (Mac OS X は UNIX)
- Windows 系　　CR + LF
- Mac (OS9 以前)　CR のみ

まとめ

文字コード

- 計算機で文字を扱うための対応表

改行コード

- OS によって使っているものが違う