

S430

文字のコード化

Character Encoding

■展示品のねらい

コンピュータの内部では、全ての文字には番号がつけられています。文字に番号をつけることを「文字をコード化する」、ある文字につけられた番号のことを「文字コード」と呼びます。この展示品では、文字コードについてゲーム感覚で学ぶことができます。



■知識プラスワン

		上1ケタ					
		20	30	40	50	60	70
F11-F16	0		0	@	P		p
	1	!	1	A	Q	a	q
	2	"	2	B	R	b	r
	3	#	3	C	S	c	s
	4	\$	4	D	T	d	t
	5	%	5	E	U	e	u
	6	&	6	F	V	f	v
	7	'	7	G	W	g	w
	8	(8	H	X	h	x
	9)	9	I	Y	i	y
	A	*	:	J	Z	j	z
	B	+	;	K	[k	[
	C	,	<	L	\	l	
	D	-	=	M]	m]
	E	.	>	N	^	n	^
	F	/	?	O	_	o	_

表1 アスキーコード (抜粋)

□US ASCIIコード

コンピュータが生まれたアメリカでは、最初は128文字分がコード化されました。これは、US ASCIIコードと呼ばれ(表1)、アルファベット大文字・小文字、数字、記号に加えて、表示することはできないがディスプレイの表示を消すとか、スピーカーからピットという音を出す、などといった機能をもたせた文字コードもあります。

□65535

日本語では、ひらがな・カタカナ・漢字を入れると数万種類も文字がありますので、65536種類の文字を表すことができるような文字コードが使われています。なぜきりのいい1万とか10万とかではなくて65536なのかというと、文字コードの表し方によるのです。コンピュータの内部では二進法で計算が行われていますが、二進法ですと大きな数を表そうとするととても長くなるため、人間にとって少しでも便利にするため十六進法が使われています。65536種類の文字を表すためには、数字として0から65535を使うということです。1から65535までで65535個。これに「0」が1個加わるので、数字としては0から65535となります。これは、16進数で表すと「0x0」から「0xFFFF」となります。十六進数2ケタで表すことができる範囲が、十進法では0から65535までなのです。これは2バイトで数を表すということになるので、こうした文字を「2バイト文字」と呼びます。

□さまざまなコード

文字コードにはいろいろな種類があって、このために時々不便なことがあります。マイクロソフト社のWindows製品で使われている文字コードは、「シフトJIS」と呼ばれています。これは「JISコード」を少し変更したものです。「UNIX」と

呼ばれるコンピュータ基本ソフトでは「EUC」が主流でした。他にも、日本語以外も扱えるようにした「Unicode」「UTF」、昔よく使われていた「EBCDIC」などなど、ものすごくたくさんあるのです。つまり、同じ文字なのに何のコードで表すのかによって、番号が変わってきてしまうということです。コンピュータは番号だけで処理しますから、「あ」という文字のつもりが、別の文字コードで処理すると全然違った文字になってしまう、という現象が起こるわけです。これを「文字化け」と言います。

□電子メール

最近では、パソコンでも携帯電話でも電子メールが使えるようになってきました。携帯電話とパソコンでは、そのしくみがかなり違っているはずですが、だいたいの場合には電子メールは文字化けせずに相手に届いています。このからくりはいったいどうなっているのでしょうか。答えは、電子メールのしくみにあります。電子メールは、本文のほかに「ヘッダ」と呼ばれる部分があります。これには、本文がどの文字コードで読まれるべきかについての指定があります。ヘッダにはこのほか、差出人、差し出し日時、受取人、題(subject)、メール配送経路記録、使ったメールソフト名などの項目もあります。中味を確実に伝えるためのこのような工夫で、わたしたちは便利に使っているのですね。

□著者 学芸員 小塩哲朗