5.4 Mã văn bản

Tuy nhiên, hầu hết máy tính không biểu diễn các ký tự ở dạng các số nhị phân thuần túy. Chúng sử dụng phiên bản mã hóa nhị phân để biểu diễn các chữ cái, các ký hiệu đặc biệt cũng như các số thập phân.

Trong ngôn ngữ tiếng Anh có 26 ký tự. Nếu chúng ta xét cả các ký tự chữ hoa, chữ thường, và các ký hiệu đặc biệt như *%+- v.v., mười chữ số thập phân, các ký tự điều khiển không in ra được như ký tự xuống dòng, v.v. thì chúng ta có 128 ký tự. Chúng ta sẽ cần 7 chữ số để biểu diễn tất cả 128 ký tự đó. Mã của các ký tự được chuẩn hóa cho phép truyền dữ liệu giữa các máy tính và các mạng. Những điều hình thành mã văn bản tiêu chuẩn.

5.4.1 ASCII

Một trong những bảng mã tiểu chuẩn phổ biến và thông nhất dụng nhất là ASCII (**A**merican **S**tandard for **I**nformation **I**nterchange – ASCII – Bảng mã tiêu chuẩn dùng để trao đổi thông tin của Mỹ - Phát âm như AS-key). ASCII sử dụng 7 bit để mã hóa 1 ký tự. Với 7 bit, ASCII có thể cung cấp 128 (2^7) sắp xếp khác nhau.

Thập phân	Bát phân	Thập lục phân	Nhị phân	Giá trị
000	000	000	00000000	NUL (Ký tự Null.)
001	001	001	00000001	SOH (Bắt đầu header – Start of Header)
002	002	002	00000010	STX (Bắt đầu văn bản Start of Text)
0004	004	004	00000100	EOT (Kết thúc truyền - End of Transmission)
005	005	005	00000101	ENQ (Enquiry)
006	006	006	00000110	ACK (Xác nhận - Acknowledgment)
007	007	007	00000111	BEL (Chuông - Bell)
008	010	008	00001000	BS (Backspace)
009	011	009	00001001	HT (Tab ngang - Horizontal Tab)
010	012	00A	00001010	LF (về đầu dòng – Line Feed)
011	013	00B	00001011	VT (Tab dọc - Vertical Tab)
012	014	00C	00001100	FF (Form Feed)
013	015	00D	00001101	CR (Xuống dòng - Carriage Return)
014	016	00E	00001110	SO (Shift Out)
015	017	00F	00001111	SI (Shift In)
016	020	010	00010000	DLE (Data Link Escape)
017	021	011	00010001	DC1 (XON) (Device Control 1)
018	022	012	00010010	DC2 (Device Control 2)
019	023	013	00010011	DC3 (XOFF)(Device Control 3)
020	024	014	00010100	DC4 (Device Control 4)
021	025	015	00010101	NAK (Từ chối xác nhận - Negative Acknowledgment)
022	026	016	00010110	SYN (Synchronous Idle)
023	027	017	00010111	ETB (End of Trans. Block)
024	030	018	00011000	CAN (Hủy bỏ - Cancel)
025	031	019	00011001	EM (End of Medium)
026	032	01A	00011010	SUB (Thay thế - Subsitute)
027	033	01B	00011011	ESC (Escape)
028	034	01C	00011100	FS (File Separator)
029	035	01D	00011101	GS (Group Separator)
030	036	01E	00011110	RS (Reqst to Send)(Rec. Sep.)
031	037	01F	00011111	US (Unit Separator)
032	040	020	00100000	SP (Khoảng trắng - Space)
033	041	021	00100001	! (dấu chấm than - exclamation mark)
034	042	022	00100010	" (dấu trích dẫn hay nháy kép - double quote)
035	043	023	00100011	# (Ký hiệu số - number sign)
036	044	024	00100100	$ (dấu dollar - dollar sign)
037	045	025	00100101	% (phần trăm - percent)

038	046	026	00100110	&	(daauss & - ampersand)
039	047	027	00100111	'	(trích dẫn đơn - single quote)
040	050	028	00101000	(	(mở ngoặc nhọn - left/open parenthesis)
041	051	029	00101001	)	(đóng ngoặc nhọn - right/closing parenth.)
042	052	02A	00101010	*	(dấu * - asterisk)
043	053	02B	00101011	+	(dấu cộng - plus)
044	054	02C	00101100	,	(dấu chám phẩy - comma)
045	055	02D	00101101	-	(dấu trự hay dấu gạch - minus or dash)
046	056	02E	00101110	(dấu chấm – dot)	
047	057	02F	00101111	/	(dấu gạch chéo - forward slash)
048	060	030	00110000	0	
049	061	031	00110001	1	
050	062	032	00110010	2	
051	063	033	00110011	3	
052	064	034	00110100	4	
053	065	035	00110101	5	
054	066	036	00110110	6	
055	067	037	00110111	7	
056	070	038	00111000	8	
057	071	039	00111001	9	
058	072	03A	00111010	:	(dấu hai chấm – colon)
059	073	03B	00111011	;	(dấu chấm phẩy - semi-colon)
060	074	03C	00111100	<	(nhỏ hơn - less than)
061	075	03D	00111101	=	(dấu bằng - equal sign)
062	076	03E	00111110	>	(lớn hơn - greater than)
063	077	03F	00111111	?	(dấu hỏi chấm – question mark)
064	100	040	01000000	@	(ký hiệu a còng - AT symbol)
065	101	041	01000001	A	
066	102	042	01000010	B	
067	103	043	01000011	C	
068	104	044	01000100	D	
069	105	045	01000101	E	
070	106	046	01000110	F	
071	107	047	01000111	G	
072	110	048	01001000	H	
073	111	049	01001001	I	
074	112	04A	01001010	J	
075	113	04B	01001011	K	
076	114	04C	01001100	L	
077	115	04D	01001101	M	
078	116	04E	01001110	N	
079	117	04F	01001111	O	
080	120	050	01010000	P	
081	121	051	01010001	Q	
082	122	052	01010010	R	
083	123	053	01010011	S	
084	124	054	01010100	T	
085	125	055	01010101	U	
086	126	056	01010110	V	
087	127	057	01010111	W	
088	130	058	01011000	X	
089	131	059	01011001	Y	
090	132	05A	01011010	Z	
091	133	05B	01011011	[	(dấu mở ngoặc vuông trái - left/opening bracket)
092	134	05C	01011100	\	(back slash)
093	135	05D	01011101	]	(dấu đóng ngoặc vuông - right/closing bracket)
094	136	05E	01011110	^	(caret/circumflex)
095	137	05F	01011111	_	(gạch dưới - underscore)
096	140	060	01100000	`	
097	141	061	01100001	a	
098	142	062	01100010	b	
099	143	063	01100011	c	
100	144	064	01100100	d	
101	145	065	01100101	e	
102	146	066	01100110	f	
103	147	067	01100111	g	
104	150	068	01101000	h	
105	151	069	01101001	i	
106	152	06A	01101010	j	
107	153	06B	01101011	k	
108	154	06C	01101100	l	
109	155	06D	01101101	m	

110	156	06E	01101110	n
111	157	06F	01101111	o
112	160	070	01110000	p
113	161	071	01110001	q
114	162	072	01110010	r
115	163	073	01110011	s
116	164	074	01110100	t
117	165	075	01110101	u
118	166	076	01110110	v
119	167	077	01110111	w
120	170	078	01111000	x
121	171	079	01111001	y
122	172	07A	01111010	z
123	173	07B	01111011	{ (mở ngoặc nhọn - left/opening brace)
124	174	07C	01111100	| (gạch đứng - vertical bar)
125	175	07D	01111101	} (đóng ngoặc nhọn - right/closing brace)
126	176	07E	01111110	~ (dấu ngã - tilde)
127	177	07F	01111111	DEL (phím delete - delete)

Bảng 5.3: Bảng ASCII

Bên cạnh mã cho các ký tự, mã cũng xác định các thông tin chẳng hạn kết thúc tệp tin, kết thúc trang, v.v… Những mã này còn gọi là các ký tự điều khiển không in được Mã ASCII được dùng để biểu diễn dữ liệu bên trong máy tính cá nhân.

5.4.2 EBCDIC

Bảng mã EBCDIC (Mã trao đổi mở rộng của số thập phân được mã hóa bằng nhị phân – phát âm là EB-si-dic) là viết tắt của **E**xtended **B**inary **C**oded **D**ecimal **I**nterchange **C**ode. EBCDIC sử dụng 8 bit để mã hóa 1 ký tự. Do đó có 256 ký tự được biểu diễn sử dụng EBCDIC. Bảng mã EBCDIC được dùng trong các máy tính lớn (mainframe) của IBM và một số máy tương tự khác.

Các mạch điện tử cũng có thể chuyển đổi các ký tự từ bảng mã ASCII sang EBCDIC và ngược lại. Chúng ta cũng có thể thực hiện việc chuyển đổi này sử dụng chương trình máy tính.

5.4.3 Unicode

Bảng mã Unicode dùng 2 byte – 16 bit- để biểu diễn mỗi chữ cái, số và ký hiệu. Với 2 byte, bảng mã Unicode có thể biểu diễn hơn 65536 các ký tự và ký hiệu khác nhau. Số này đủ để biểu diễn hết mỗi ký tự, ký hiệu trên thế giới, như tiếng Trung, tiếng Hàn Quốc, tiếng Nhật, tiếng Việt, tiếng Thái Lan, v.v. Bộ các ký tự được tìm thấy trong các văn bản cổ. Ưu điểm chính mà Unicode hơn hẳn so với các bảng mã khác là nó có khả năng tương thích với bảng mã ASCII. 256 mã đầu tiên trong Unicode là 256 mã được dùng trong bảng mã ASCII. Unicode sau này còn mở rộng hơn rất nhiều so với tập các ký tự ASCII chuẩn.

Một trong những kỹ thuật Unicode quan trọng nhất là UTF-8 (**Unicode Transformation Format** - Định dạng chuyển đổi Unicode 8 bit) trong đó sử dụng các ký tự mã hóa Unicode với chiều dài biến đổi. Không giống cách mã hóa Unicode thông thường; trong UTF-8, các mã có độ dài khác nhau được dùng để mã hóa bộ ký tự (tập hợp các ký hiệu). UTF-8 mã hóa 128 ký tự trong bộ bý tự Unicode (tương ứng với ASCII) sử dụng một **single octet** (các nhóm 8 bit) với cùng giá trị nhị phân trong ASCII. UTF-8 tương thích với ASCII theo các đặc điểm này. Với lý do này, UTF-8 trở thành bộ ký tự mã hóa chủ chốt cho các tệp tin, trang web và phần mềm làm việc với dữ liệu dạng văn bản.

Trước khi chúng ta đi sâu và hiểu rõ hơn về xử lý dữ liệu, chúng ta cần nhìn sâu vào hai thành phần xử lý trong máy tính : **CPU** và **bộ nhớ.**