Home: WWW Information: Publishing Web Docs:ASCII Characters

ASCII Characters

Operating systems, programming/scripting lanuages, protocols and text processing systems use characters in different ways. This summarizes the character set and some of the special uses of and restrictions on characters.

The ASCII (7-bit) (American National Standard Code for Information Interchange) code set is defined in ANSI Spec X3.4. Extended (8-bit codes), as defined in ISO8859-1, (Latin 1) can also be used in HTML.

- Control Characters
- Printable Characters
- Usage of Special Characters
- Special Characters allowed in names and addresses
- ISO Latin and extended ASCII Character References

Text data: ASCII

See also: Special Character Names
Character Usage

There are two main codes in use for character data: ASCII and EBCDIC. EBCDIC is used almost exclusively on IBM machines and their clones. On most other computer systems, , ASCII is used, so that is all we will discuss here. ASCII is by far the more common of the two.

ASCII stands for American Standard Code for Information Interchange. It contains a binary code for all the characters generated by the keyboard, and a few others that are not generated by all keyboards.

The standard ASCII set consists of 128 binary codes, from 000 0000 to 111 1111. The msb of the byte is not written because it is sometimes reserved for a parity bit (an error check: see later) and on some micro computers another 128 special symbols (graphic characters or mathematical symbols) are defined using this eighth bit. Since its use varies from one system to another, we will explicitly write only the first 7 bits.

HTML Character References use the Decimal code. e.g. &#64; = '@' .
URL Encoding uses Hex characters (e.g. %40 = @)

Control Characters

				CTRL	(^D means to hold the CTRL key and hit d)
Oct	Dec	Char	Hex	Key	Comments
\000	0	NUL	\x00	^@ \0	(Null byte)
\001	1	SOH	\x01	^A	(Start of heading)
\002	2	STX	\x02	^B	(Start of text)
\003	3	ETX	\x03	^C	(End of text) (see: UNIX keyboard CTRL)
\004	4	EOT	\x04	^D	(End of transmission) (see: UNIX keyboard CTRL)
\005	5	ENQ	\x05	^E	(Enquiry)
\006	6	ACK	\x06	^F	(Acknowledge)
\007	7	BEL	\x07	^G	(Ring terminal bell)
\010	8	BS	\x08	^H \b	(Backspace) (\b matches backspace inside [] only) (see: UNIX keyboard CTRL)
\011	9	HT	\x09	^I \t	(Horizontal tab)
\012	10	LF	\x0A	^J \n	(Line feed) (Default UNIX NL) (see End of Line below)
\013	11	VT	\x0B	^K	(Vertical tab)
\014	12	FF	\x0C	^L \f	(Form feed)
\015	13	CR	\x0D	^M \r	(Carriage return) (see: End of Line below)
\016	14	SO	\x0E	^N	(Shift out)
\017	15	SI	\x0F	^O	(Shift in)
\020	16	DLE	\x10	^P	(Data link escape)
\021	17	DC1	\x11	^Q	(Device control 1) (XON) (Default UNIX START char.)
\022	18	DC2	\x12	^R	(Device control 2)
\023	19	DC3	\x13	^S	(Device control 3) (XOFF) (Default UNIX STOP char.)
\024	20	DC4	\x14	^T	(Device control 4)
\025	21	NAK	\x15	^U	(Negative acknowledge) (see: UNIX keyboard CTRL)
\026	22	SYN	\x16	^V	(Synchronous idle)
\027	23	ETB	\x17	^W	(End of transmission block)
\030	24	CAN	\x18	^X	(Cancel)
\031	25	EM	\x19	^Y	(End of medium)
\032	26	SUB	\x1A	^Z	(Substitute character)
\033	27	ESC	\x1B	^[	(Escape)
\034	28	FS	\x1C	^\	(File separator, Information separator four)
\035	29	GS	\x1D	^]	(Group separator, Information separator three)
\036	30	RS	\x1E	^^	(Record separator, Information separator two)
\037	31	US	\x1F	^_	(Unit separator, Information separator one)
\177	127	DEL	\x7F	^?	(Delete) (see: UNIX keyboard CTRL)

Printable Characters

Specials (32-47)

(See: Special Character Names)

\040	32	" "	\x20	(space)
\041	33	!	\x21	EXCLAMATION POINT(bang)
\042	34	"	\x22	QUOTATION MARK, DIAERESIS
\043	35	#	\x23:	NUMBER SIGN (Pound sign) (see: UNIX keyboard CTRL)
\044	36	$	\x24	DOLLAR SIGN
\045	37	%	\x25	PERCENT SIGN
\046	38	&	\x26	AMPERSAND
\047	39	'	\x27	APOSTROPHE, RIGHT SINGLE QUOTATION MARK, ACUTE ACCENT (single quote)
\050	40	(	\x28	LEFT PARENTHESIS (open parenthesis)
\051	41	)	\x29	RIGHT PARENTHESIS (close parenthesis)
\052	42	*	\x2A	ASTERISK
\053	43	+	\x2B	PLUS SIGN
\054	44	,	\x2C	COMMA, CEDILLA
\055	45	-	\x2D	HYPHEN, MINUS SIGN
\056	46	.	\x2E	PERIOD, DECIMAL POINT, (Full Stop)
\057	47	/	\x2F	SLANT (SOLIDUS), slash

Digits

\060	48	0	\x30
\061	49	1	\x31
\062	50	2	\x32
\063	51	3	\x33
\064	52	4	\x34
\065	53	5	\x35
\066	54	6	\x36
\067	55	7	\x37
\070	56	8	\x38
\071	57	9	\x39

Specials (58-64)

\072	58	:	\x3A	COLON
\073	59	;	\x3B	SEMICOLON
\074	60	<	\x3C	LESS-THAN SIGN (left angle bracket)
\075	61	=	\x3D	EQUALS SIGN
\076	62	>	\x3E	GREATER-THAN SIGN (right angle bracket)
\077	63	?	\x3F	QUESTION MARK
\100	64	@	\x40	COMMERCIAL AT † (see: UNIX keyboard CTRL)

Latin Capital Letters

\101	65	A	\x41
\102	66	B	\x42
\103	67	C	\x43
\104	68	D	\x44
\105	69	E	\x45
\106	70	F	\x46
\107	71	G	\x47
\110	72	H	\x48
\111	73	I	\x49
\112	74	J	\x4A
\113	75	K	\x4B
\114	76	L	\x4C
\115	77	M	\x4D
\116	78	N	\x4E
\117	79	O	\x4F
\120	80	P	\x50
\121	81	Q	\x51
\122	82	R	\x52
\123	83	S	\x53
\124	84	T	\x54
\125	85	U	\x55
\126	86	V	\x56
\127	87	W	\x57
\130	88	X	\x58
\131	89	Y	\x59
\132	90	Z	\x5A

Specials (91-96)

\133	91	[	\x5B	LEFT (SQUARE) BRACKET (open bracket) †
\134	92	\	\x5C	REVERSE SLANT (REVERSE SOLIDUS) (backslash, backslant) †
\135	93	]	\x5D	RIGHT (SQUARE) BRACKET (closing bracket) †
\136	94	^	\x5E	CIRCUMFLEX ACCENT †
\137	95	_	\x5F	UNDERLINE (LOW LINE)
\140	96	`	\x60	LEFT SINGLE QUOTATION MARK, GRAVE ACCENT †

Latin Small Letters

\141	97	a	\x61
\142	98	b	\x62
\143	99	c	\x63
\144	100	d	\x64
\145	101	e	\x65
\146	102	f	\x66
\147	103	g	\x67
\150	104	h	\x68
\151	105	i	\x69
\152	106	j	\x6A
\153	107	k	\x6B
\154	108	l	\x6C
\155	109	m	\x6D
\156	110	n	\x6E
\157	111	o	\x6F
\160	112	p	\x70
\161	113	q	\x71
\162	114	r	\x72
\163	115	s	\x73
\164	116	t	\x74
\165	117	u	\x75
\166	118	v	\x76
\167	119	w	\x77
\170	120	x	\x78
\171	121	y	\x79
\172	122	z	\x7A

Specials (123-126)

```
\173 123  {  \x7B  LEFT BRACE (LEFT CURLY BRACKET) (open brace) †
\174 124  |  \x7C  VERTICAL LINE (pipe) †
\175 125  }  \x7D  RIGHT BRACE (RIGHT CURLY BRACKET) (closing brace) †
\176 126  ~  \x7E  TILDE (OVERLINE) (squiggle) †
```

Control (127)

```
\177 127 DEL \x7F ^?            (Delete)  (see: UNIX keyboard CTRL)

 † The characters following the letters may be used for additional
letters in countries with alphabets containing more than 26 letters.These characters should not bae used in international interchange
without determining that there is agreement between sender and recipient.
```

Usage of Special Characters

End of Line character

```
 End of Line varies depending on the operating system:
  DOS/Windows:  <CR><LF>
  Macintosh:... <CR>
  UNIX..........<LF>  (See File Format Notes for more information.)
```

UNIX Keyboard Control Characters

```
:
  The default keyboard control characters vary depending on the UNIX system.
  Most people change them with the stty command in their .profile.
                           SysV  Sun/Solaris  HP/UX
  Erase (character delete)  #      <DEL>       <BS> (^H)
  Kill (line delete)        @       ^U          @
  Intr (Interupt process) <DEL>     ^C        <DEL>
  EOF  (End of File)       ^D       ^D         ^D
                          EOF Signals End of File for characters input from
                          the terminal.  Also causes shell to terminate.
```

Special Characters allowed in names and addresses:

```
 Note: The only characters other than letters and digits which appear to
       be universly acceptable are - (dash) and _ (underscore) and you
       have to watch out for '-' which can be interpreted as minus when
       used in a name in certain perl scripts.

         (1)      (2) (3)
Octal   UNIX DOS SMTP URL (HTML - allows all but <, >, &,and  ")
\011 TAB
\040 " "          -     Spaces can be used in mail addresses if the addr. is quoted.
\041  !       *   *   *     ! can cause problems in csh in UNIX.
\042  "
\043  #   *   *   *      (see: UNIX keyboard CTRL)
\044  $       *   *   *
\045  %   *   *   *
\046  &       *   *
\047  '       *   *   *
\050  (       *
\051  )       *
\052  *           *   *
\053  +   *       *   *   (URL's sometimes use + for space)
\054  ,   *
\055  -   *   *   *   *
\056  .   *
\057  /           *
\072  :   *
\073  ;
\074  <
\075  =   *       *
\076  >
\077  ?           *
\100  @   *   *         (see: UNIX keyboard CTRL)
\133  [
\134  \
\135  ]
\136  ^       *   *
\137  _   *   *   *   *
\140  `       *   *
\173  {       *   *
\174  |           *
\175  }       *   *
\176  ~   *   *   *
```

(1) UNIX - Any character except "/" (slash) is allowed
in a UNIX file name but many are not recommended
because they cause problems in scripting and/or
programming languaages dealing with the files.
(2) SMTP - (Simple Mail Transfer Protocol)
(3)URI/URL - Uniform Resource Identifier/Locator. Other characters can
be used but require encoding with % and the HEX value (e.g. @ = %40)
(Space is sometimes encoded as "+".)
(4) HTML - HyperText Markup Language requires 4 ASCII characters to be
encoded as character or entity references (escape sequences).

ASCII characters with special meaning in HTML so they must be encoded:

Character	Character Reference	Entity Reference
<	&#60;	&lt;
>	&#62;	&gt;
&	&#38;	&amp;
"	&#34;	&quot;

Other common non-ASCII character encodings for HTML:

Description	**Code**	**Entity name**	**Octal Code**
e, acute accent	&#233; --> é	&eacute; --> é	\351 (octal) = é
ampersand	&#38; --> &	&amp; --> &	
registered trademark	&#174; --> ®	&reg; --> ®	
copyright	&#169; --> ©	&copy; --> ©	
trademark	&#153; --> ™	<SUP><FONT SIZE=-1>TM</FONT></SUP> --> TM	

Other HTML Character Reference Tables

ISO8859-1, (Latin 1) notes and Character List at Best Business Solutions (BBS).
Extended ASCII (same as ISO859-1) at emory.edu

ISO (International Organization for Standardization) defines several character sets.
e.g. the ISO 8859 series.
HTML Character Entity names are defined targnet.org and uni-passau.

IBM
IBM uses (EBCDIC) Extended Binary Coded Decimal Interchange Code
(8-bit) coding on most of their systems.
They uses code pages to specify charact sets for keyboards, displays,
printers, ... for DOS, AIX, Mainframes, ...
Standard DOS code pages are:

437	United States
850	Multilingual (Latin 1)
852	Slavic (Latin 2)
863	Canadian-French
865	Nordic (Norwegian, Danish)
860	Portuguese

See:
IBM OS/390 Code Pages

General Info. on Code Pages

See also: BYTE article 'Organizing Babylon' on international character sets.

Netscape Character Sets
MIME Charset parameter in HTTP. If the server includes this parameter in its
response, Netscape Navigator will change its character set appropriately.
For example:

```
Content-Type: text/html;charset=iso-8859-1
Content-Type: text/html;charset=iso-2022-jp
```

The charset names recognized by Netscape Navigator 1.1 are specified in
RFC 1700 (except for the names that begin with "x-".) These include:

```
us-ascii
iso-8859-1
iso-2022-jp
x-sjis
x-euc-jp
x-mac-roman
```

Additionally, the following aliases are recognized for us-ascii:

```
ansi_x3.4-1968
iso-ir-6
ansi_x3.4-1986
iso_646.irv:1991
ascii
iso646-us
us
ibm367
cp367
```

Return to Publishing

Send comments and suggestions to mcbride@cc.bellcore.com