Bit by Bit: The Development of Character Encoding Standards

Aditya Singhvi

Honors ATCS: Programming Languages

Dr. Eric Nelson

September 10, 2020

Word Count: 1226

Character encoding refers to the representation of textual characters as numbers to facilitate the storage and transmission of text-based data.[1] Traditional computers store data using bits, binary units that can either be “on” or “off,” with each character represented using a unique bit-sequence. In order to ensure character encoding compatibility across machines, a plethora of standards — such as EBCDIC, ASCII, and Unicode — have been developed over the years. Forced to a uniform standard with the advent of web-based technologies, modern computing applications generally comply to the Unicode standard, maintained by the Unicode Consortium and compatible with the older ASCII standard.[2]

One of the earliest encoding standards — Morse code — arose as electronic communications grew in the 19th century with the invention of the telegraph. Invented by Samuel Morse in the 1830s, Morse code uses a variable-length binary encoding system — traditionally represented as dots and dashes — to transmit alphanumeric characters.[3] Yet, as the nascent computing industry advanced, the limitations of Morse code forced computer scientists such as Jean-Maurice-Émile Baudot and Herman Hollerith to invent more complex encoding systems.[4] Hollerith devised the concept of a punch card (See *Figure 1*) to tabulate data using a binary encoding system, helping the U.S. government save approximately five million dollars in carrying out the 1890 census.[5] Emboldened by his success, he founded the Tabulating Machine Company in 1896, which merged to form the International Business Machine Corporation (IBM) in 1924.[6]

IBM quickly rose to the forefront of computing technology with the introduction of the IBM Card in 1928, which quickly supplanted older punch cards with its 80-column, 12-row design (as seen in *Figure 1*) that could store nearly twice as much data as the contemporary

[1] Christensson, Per. "Character Encoding Definition." TechTerms. (September 24, 2010). https://techterms.com/definition/characterencoding.

[2] Christensson, 2010.

[3] Britannica School, s.v. "Morse Code," accessed September 7, 2020, https://school-eb-com.puffin.harker.org/levels/high/article/Morse-Code/53835.

[4] Britannica School, s.v. "Jean-Maurice-Émile Baudot," accessed September 8, 2020, https://school-eb-com.puffin.harker.org/levels/high/article/Jean-Maurice-%C3%89mile-Baudot/13803.

[5] Frank da Cruz, "Herman Hollerith," Columbia University Computing History, last modified July 2020, accessed September 8, 2020, http://www.columbia.edu/cu/computinghistory/hollerith.html.

[6] da Cruz, "Herman Hollerith."

45-column standard.[7] With the introduction of machines such as the IBM 700 series in the 1950s, this punch card encoding standard was adapted to create the six-bit, 48-character Binary-Coded Decimal Interchange Code (BCDIC), with the two high bits representing the two "zone rows" of the IBM card and four low bits representing the ten "digit rows."[8]

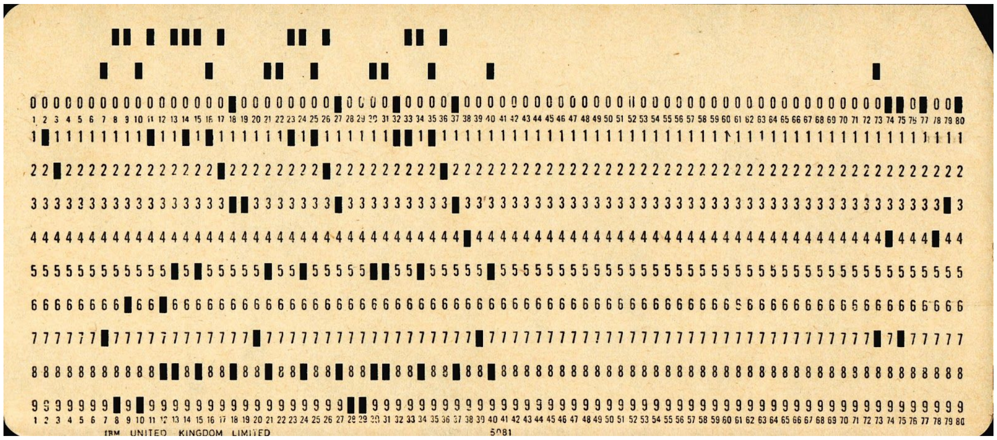


Figure 1. A 12-row, 80-column IBM punch card from the 1940s. Each byte was encoded on a single column, with the two "zone rows" at the top of the card being punched to indicate alphabetic coding.[9,10]

As the field of computing expanded, the six-bit BCDIC architecture, which allowed for 64 unique character encodings, began to prove insufficient for certain data processing applications.[11] Thus, with the introduction of the System/360 line of computers in 1963, IBM developed Extended BCDIC (EBCDIC), which allowed for 256 unique characters with an eight-bit architecture.[12] Although IBM initially planned to support both ASCII (the standard for most other machines at the time) and EBCDIC, a series of miscommunications within the company led to the rushed rollout of the System/360 line without ASCII support.[13]

[7] Emerson W. Pugh and Lars Heide, "STARS:Punched Card Equipment," IEEE Global History Network, last modified May 11, 2011, accessed September 8, 2020, https://web.archive.org/web/20120511034402/http://www.ieeeghn.org/wiki/index.php/STARS%3APunched_Card_Equipment.

[8] Wikipedia Contributors, "Character Encoding," Wikipedia, last modified August 25, 2020, accessed September 8, 2020, https://en.wikipedia.org/wiki/Character_encoding.

[9] Pete Birkinshaw, *A 12-row/80-column IBM punched card from the mid-twentieth century*, photograph, Wikipedia, November 7, 2020, accessed September 8, 2020, https://en.wikipedia.org/wiki/Punched_card#/media/File:Used_Punchcard_(5151286161).jpg.

[10] Emerson W. Pugh and Lars Heide, "STARS:Punched Card Equipment."

[11] Charles E. Mackenzie, *Coded Character Sets, History and Development* (Addison-Wesley Publishing Company, 1980), 120, PDF.

[12] Mackenzie, 121.

[13] Wikipedia Contributors, "Character Encoding."

The EBCDIC code table is divided into quadrants based on the four most significant bits (See *Figure 2*).[14] The first quadrant is reserved for control sequences, the second for special graphics, and the third and fourth for alphanumeric characters.[15] Although this structure solved the problem of insufficient space, the punch-card-friendly design of EBCDIC presented numerous challenges for programmers. As shown in *Figure 2*, the letter encodings are not consecutive, with the gaps overly complicating theoretically simple programs.[16] Furthermore, sorting a sequence encoded with EBCDIC led to lowercase letters being placed before uppercase letters, which were placed before numerals — the opposite of the ASCII collating standard.[17] Thus, because of these issues, modern computer systems have largely moved away from EBCDIC and toward ASCII-based character encoding standards.[18]

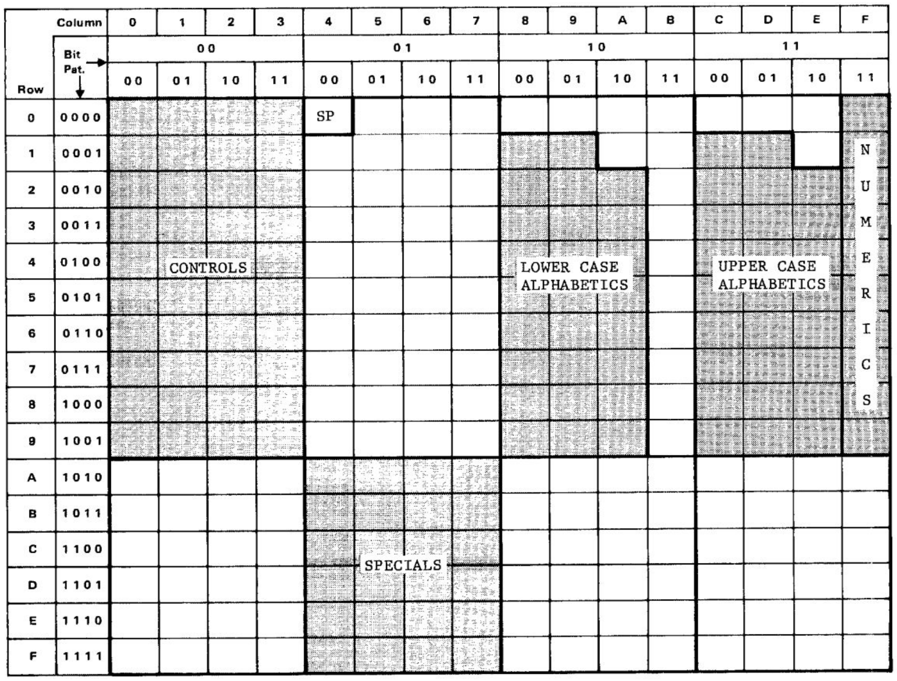


Figure 2. A code table laying out the general structure of eight-bit EBCDIC, with columns representing the four high-order bits and rows representing the low-order bits. The empty spaces from Row A to Row F within the alphabetical characters caused problems for programmers.[19]

[14] Mackenzie, *Coded Character*, 125.
[15] Mackenzie, 139.
[16] Wikipedia Contributors, "EBCDIC," Wikipedia, last modified August 26, 2020, accessed September 8, 2020, https://en.wikipedia.org/wiki/EBCDIC.
[17] Wikipedia Contributors, "EBCDIC."
[18] Wikipedia Contributors, "Character Encoding."
[19] Mackenzie, *Coded Character*, 140.

The American Standard Code for Information Interchange, or ASCII, was developed beginning in the late 1950s as the need for a unifying encoding standard became more apparent.[20] After cataloging over sixty different methods of representation for just the Latin alphabet, IBM Computer scientist Bob Bemer, now termed the "father of ASCII," began working with the American Standards Association (ASA) to help create such a standard.[21] The ASA formed Committee X3 in 1960, bringing computer scientists and statisticians together to achieve this goal.[22] After reviewing requirements across industries, the committee decided on a seven-bit architecture capable of encoding 128 unique sequences, publishing the first edition of ASCII in 1963 and revising it significantly in 1967.[23]

The 1967 ASCII code table, shown in *Figure 3*, is divided into eight 16-character columns ordered by the three high-priority bits, with the first two columns dedicated to controls and the remaining six reserved for graphics. The Space character, denoted by SP in *Figure 3*, was determined to be a graphic as opposed to a control after much debate within the committee, and was thus placed in the third column.[24] Unlike EBCDIC, ASCII ensures consecutive bit patterns within the alphabet, with corresponding uppercase and lowercase letter encodings differing by a single bit. Furthermore, ASCII collating order places numerals first, followed by uppercase and then lowercase letters, the opposite of EBCDIC. In 1968, U.S. President Lyndon B. Johnson adopted ASCII as a federal standard, mandating that all government computers bought after 1969 be ASCII-compatible;[25] most modern text editors continue to support ASCII, albeit supporting a much larger character set as well with the Unicode standard.[26]

[20] Mackenzie, 211.

[21] Bemer, Bob. "A Story of ASCII." Last modified November 8, 2002. PDF.

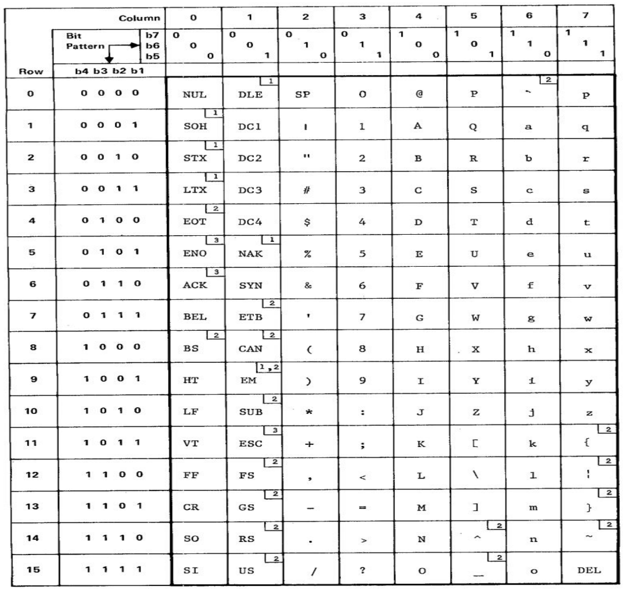
[22] Bemer, Bob. "A Story of ASCII."

[23] Mackenzie, *Coded Character*, 216.

[24] Mackenzie, 222.

[25] Lyndon B. Johnson to Heads of Departments and Agencies, memorandum, "Memorandum Approving the Adoption by the Federal Government of a Standard Code for Information Interchange," March 1968, accessed September 9, 2020, https://www.presidency.ucsb.edu/documents/memorandum-approving-the-adoption-the-federal-government-standard-code-for-information.

[26] McClure, Wanda L., and Stan A. Hannah. "Communicating globally: the advent of Unicode." Computers in Libraries, May 1995, 19+. Gale General OneFile (accessed September 9, 2020). https://link-gale-com.puffin.harker.org/apps/doc/A17155100/ITOF?u=harker&sid=ITOF&xid=2b8b2bd2.



	Column	0	1	2	3	4	5	6	7
	Bit Pattern b7	0	0	0	0	1	1	1	1
	b6	0	0	1	1	0	0	1	1
	b5	0	1	0	1	0	1	0	1
Row	b4 b3 b2 b1								
0	0 0 0 0	NUL	DLE [1]	SP	0	@	P	` [2]	p
1	0 0 0 1	SOH [1]	DC1	!	1	A	Q	a	q
2	0 0 1 0	STX [1]	DC2	"	2	B	R	b	r
3	0 0 1 1	LTX [1]	DC3	#	3	C	S	c	s
4	0 1 0 0	EOT [2]	DC4	$	4	D	T	d	t
5	0 1 0 1	ENO [3]	NAK [1]	%	5	E	U	e	u
6	0 1 1 0	ACK [3]	SYN	&	6	F	V	f	v
7	0 1 1 1	BEL	ETB [2]	'	7	G	W	g	w
8	1 0 0 0	BS [2]	CAN [2]	(	8	H	X	h	x
9	1 0 0 1	HT	EM [1,2]	)	9	I	Y	i	y
10	1 0 1 0	LF	SUB [2]	*	:	J	Z	j	z
11	1 0 1 1	VT	ESC [3]	+	;	K	[	k	{ [2]
12	1 1 0 0	FF	FS [2]	,	<	L	\	l	¦ [2]
13	1 1 0 1	CR	GS [2]	–	=	M	]	m	} [2]
14	1 1 1 0	SO	RS [2]	.	>	N	^ [2]	n	~ [2]
15	1 1 1 1	SI	US [2]	/	?	O	_ [2]	o	DEL

Figure 3. The 1967 ASCII code table, with columns representing the three high-order bits of the seven-bit architecture. The first two columns represent control characters, while the next six represent graphics. Digits collate low to letters, partially to ensure a one-bit difference between lower and uppercase letters.[27]

In the 1980s, an important limitation of ASCII began to emerge: with 128 characters, it had sufficient space only to encode a Latin-based alphabet.[28] Three engineers — Joseph Becker, Lee Collins, and Mark Davis — began investigating the problem in 1988.[29] Becker's original Unicode proposal formed the basis for the version launched in 1991, structured as a sixteen-bit architecture capable of supporting 65,536 unique encoding combinations.[30] Becker ensured that the first 128 Unicode bit sequences — beginning with a series of nine zeroes — corresponded directly to ASCII, making Unicode back-compatible with the existing standard.[31] With the launch of Unicode 2.0 in 1996, Unicode was expanded to its modern 21-bit space capable of encoding a much larger variety of ancient and modern scripts.[32] Unicode characters are encoded

[27] Mackenzie, 247.
[28] Unicode Inc., "History of Unicode," Unicode, last modified August 31, 2006, accessed September 9, 2020, https://www.unicode.org/history/summary.html.
[29] Unicode Inc., "History of Unicode."
[30] Joseph D. Becker, Unicode 88, 4, August 29, 1988, accessed September 9, 2020, https://unicode.org/history/unicode88.pdf.
[31] Becker, Unicode 88, 5.
[32] Unicode Inc., "Frequently Asked Questions: UTF-8, UTF-16, UTF-32 & BOM," Unicode, https://www.unicode.org/faq/utf_bom.html.

between hexadecimal addresses *x0000* and *x10FFFF*, with seventeen distinct “planes” referenced by the first two digits of the address that correspond to the highest-priority byte.[33] The vast majority of commonly-used characters fall within the first plane of 65,536 characters, called the Basic Multilingual Plane (BMP).[34] Despite its standardization, Unicode data can be represented using several distinct Unicode Transformation Formats (UTFs) including UTF-8, UTF-16, and UTF-32, with the identifying numbers corresponding to the size of a code unit in that encoding.[35] For instance, UTF-8 is a variable-length encoding system that encodes in eight-bit units, with up to four units used to represent a single character.[36] Because of its compatibility with ASCII, UTF-8 is the most common encoding format across web-based applications to transfer data.[37] UTF-16, used in Java and Windows, encodes characters with up to two sixteen-bit units, although BMP characters (the most frequently used) each require a single unit.[38] UTF-32, although inefficiently stored, presents the advantage of fixed-length encoding that ensures constant-time access to each code point as opposed to the linear-time access of UTF-8 and UTF-16.[39]

Unicode has flourished since its launch in 1991, becoming the global standard for character encoding with over 95% of websites encoding in UTF-8.[40] The Unicode Standard, maintained by the non-profit Unicode Consortium, currently encompasses 143,859 unique characters across 154 scripts as of Version 13.0.0, which was launched in March 2020.[41] Although other standards (such as ISO-8859 and Windows-1251) remain entrenched in older systems, nearly all newly developed systems are either Unicode or ASCII-compliant.[42] With an expandable, comprehensive, and continually maintained architecture, Unicode seems poised to remain the character encoding standard in traditional computers for the foreseeable future.

[33] Unicode Inc., "The Unicode® Standard: A Technical Introduction," Unicode, last modified August 2019, accessed September 9, 2020, https://www.unicode.org/standard/principles.html.
[34] Unicode Inc., "The Unicode®."
[35] Unicode Inc., "Frequently Asked," Unicode.
[36] Unicode Inc., "Frequently Asked."
[37] Unicode Inc., "The Unicode®," Unicode.
[38] Unicode Inc., "The Unicode®."
[39] Unicode Inc., "The Unicode®."
[40] W3 Techs, "Usage of character encodings broken down by ranking," Web Technology Surveys, accessed September 9, 2020, https://w3techs.com/technologies/cross/character_encoding/ranking.
[41] Unicode Inc., "Unicode® 13.0.0," Unicode, last modified March 2020, accessed September 9, 2020, http://www.unicode.org/versions/Unicode13.0.0/.
[42] W3 Techs, "Usage of character.”

Bibliography

Becker, Joseph D. *Unicode 88*. August 29, 1988. Accessed September 9, 2020. https://unicode.org/history/unicode88.pdf.

Bemer, Bob. "A Story of ASCII." Last modified November 8, 2002. PDF.

Birkinshaw, Pete. *A 12-row/80-column IBM punched card from the mid-twentieth century*. Photograph. Wikipedia. November 7, 2020. Accessed September 8, 2020. https://en.wikipedia.org/wiki/Punched_card#/media/File:Used_Punchcard_(5151286161).jpg.

Britannica School, s.v. "Jean-Maurice-Émile Baudot," accessed September 8, 2020, https://school-eb-com.puffin.harker.org/levels/high/article/Jean-Maurice-%C3%89mile-Baudot/13803.

———. "Morse Code," accessed September 7, 2020, https://school-eb-com.puffin.harker.org/levels/high/article/Morse-Code/53835.

Christensson, Per. "Character Encoding Definition." TechTerms. (September 24, 2010). Accessed September 9, 2020. https://techterms.com/definition/characterencoding.

da Cruz, Frank. "Herman Hollerith." Columbia University Computing History. Last modified July 2020. Accessed September 8, 2020. http://www.columbia.edu/cu/computinghistory/hollerith.html.

Johnson, Lyndon B. Memorandum to Heads of Departments and Agencies, memorandum, "Memorandum Approving the Adoption by the Federal Government of a Standard Code for Information Interchange," March 1968. Accessed September 9, 2020. https://www.presidency.ucsb.edu/documents/memorandum-approving-the-adoption-the-federal-government-standard-code-for-information.

Mackenzie, Charles E. *Coded Character Sets, History and Development*. Addison-Wesley Publishing Company, 1980. PDF.

McClure, Wanda L., and Stan A. Hannah. "Communicating globally: the advent of Unicode." *Computers in Libraries*, May 1995, 19+. *Gale General OneFile* (accessed September 9, 2020). https://link-gale-com.puffin.harker.org/apps/doc/A17155100/ITOF?u=harker&sid=ITOF&xid=2b8b2bd2.

Pugh, Emerson W., and Lars Heide. "STARS: Punched Card Equipment." IEEE Global History Network. Last modified May 11, 2011. Accessed September 8, 2020. https://web.archive.org/web/20120511034402/http://www.ieeeghn.org/wiki/index.php/STARS%3APunched_Card_Equipment.

Unicode Inc. "Frequently Asked Questions: UTF-8, UTF-16, UTF-32 & BOM." Unicode. https://www.unicode.org/faq/utf_bom.html.

———. "History of Unicode." Unicode. Last modified August 31, 2006. Accessed September 9, 2020. https://www.unicode.org/history/summary.html.

———. "Unicode® 13.0.0." Unicode. Last modified March 2020. Accessed September 9, 2020. http://www.unicode.org/versions/Unicode13.0.0/.

———. "The Unicode® Standard: A Technical Introduction." Unicode. Last modified August 2019. Accessed September 9, 2020. https://www.unicode.org/standard/principles.html.

W3 Techs. "Usage of character encodings broken down by ranking." Web Technology Surveys. Accessed September 9, 2020. https://w3techs.com/technologies/cross/character_encoding/ranking.

Wikipedia Contributors. "Character Encoding." Wikipedia. Last modified August 25, 2020. Accessed September 8, 2020. https://en.wikipedia.org/wiki/Character_encoding.

———. "EBCDIC." Wikipedia. Last modified August 26, 2020. Accessed September 8, 2020. https://en.wikipedia.org/wiki/EBCDIC.