Standards

EDITOR'S NOTE

This paper describes a tentative standard code which has been proposed to the membership of the ASA X3.2 Subcommittee on Character Sets and Input/Output Media. The X3.2 Subcommittee is now examining this work in detail and, in my opinion, will probably recommend as an American Standard a code which is quite similar in principle to the one described here although it may differ in some details. Because of the urgency of this work and the very great amount of effort which has already gone into the development of this proposal, it is presented here, informally, for the information and scrutiny of the ACM membership although at proof time it had not yet been reviewed by the ACM Standards Committee for consideration as an ACM Standard. The authors and this Department will welcome your comments.

*Design of an Improved** Transmission/Data Processing Code

R. W. BEMER, H. J. SMITH, JR., F. A. WILLIAMS, JR.
IBM Corp., White Plains, N. Y.

Historically there has been strong difference of opinion in the construction of 6-bit (64-character) data codes, based upon whether the code is to be used for communications or data processing. This paper reports on investigation of an improved code which meets transmission requirements and requires very little modification for varied data processing usage.

It has been evident from the workings of the ASA Subcommittee X3.2 that the transmission people are not as adaptable to modifications as the data processing people. This is simply a matter of inflexibility of existing semi-mechanical communications equipment compared to the general-purpose nature of electronic data processing equipment.

The major obstacle lies in the collating, or ranking, sequence of the characters of the set. It is true that a large proportion of the ordered files of today are sequenced on numeric keys alone. However, a substantial proportion of these files are ordered on keys which contain alphabetic and special punctuation characters. If a standard code changes the relative ranking of such characters the presently ordered files will all have to be fully reordered to the new sequence, a process requiring a great expenditure of machine time. Transformation of one bit representation to another is relatively simple when the sequencing property is ignored. However, one should try to guarantee that the files are still in proper order after such conversion [1].

* Revision, 15 Mar. 1961.

There are three inputs to the collating problem:

(1) The most prevalent ranking in the U.S. is that established by IBM equipment, particularly the 705. In order are the blank, special characters, the alphabet, the digits. The critical point here is that the digits are higher than the alphabet, for whatever reason. The United Kingdom and certain other U.S. manufacturers (Sperry Rand and RCA) rank the digits lower than the alphabet.

(2) The desire of communications people, as first evidenced by Fieldata [2, 3, 4], is to have the 6-bit set collapsible to a 5-bit Baudot-type set with effectively the same characters. This is to utilize existing Baudot-Teletype equipment with simple modification.

(3) Certain punctuation characters, by universally accepted practice, should collate low to alphabets, digits, and other special characters. For example, the following two names would normally be ordered:

Roberts, A. B.
Robertson, X.

whereas the Fieldata code, because the comma ranks higher than the alphabet, would yield an ordering:

Robertson, X.
Roberts, A. B.

Expansion and contraction between any of the 4-, 5-, 6-, 7- and 8-bit code sets demand a certain uniformity and simplicity. Thus the alphabet should be reserved to two contiguous quadrants of the four quadrants of the 6-bit set. The choice now appears as in Figure A.

(typified by)	Quadrant 1	2	3	4
Bendix G-20, GAMMA 60	Alphabet	Alphabet	Digits	Special
Fieldata	Alphabet	Alphabet	Special	Digits, special
IBM Stretch	Blank, special	Alphabet	Alphabet	Digits
U.K. [5]	Blank, special	Digits	Alphabet	Alphabet

FIG. A

In the opinion of the authors neither the Fieldata code nor the U.K. code meet the criterion for 5-bit Baudot-like operation completely, even though that was one of the major design requirements. A Baudot type of code is formed essentially as follows:

Letters	Control
Digits and special	

32 5-bit combinations

In any 2-mode code for paper tape, three of the control codes, DELETE, FIGURE SHIFT and LETTER SHIFT, *must* invariably be common to both shifts. DELETE must be all 1's (all punched on paper tape) and MASTER SPACE must be all 0's (unpunched tape). MASTER SPACE, BLANK, and ESCAPE preferably appear in both shifts. Such controls as LINE FEED, CARRIAGE RETURN need appear in only one shift, but operation is more complicated.

Some of these functions may be combined in a single code combination. DELETE/LETTER SHIFT is a single code in Baudot. FIGURE SHIFT is synonymous with one of the three functions possible to ESCAPE. [6]

Since DL, FS and LS must be common to both modes, Fieldata loses the Y and Z of the alphabet and the − and + characters in the collapsed 5-bit mode. This is not tolerable because some words are spelled using Y and Z. Similarly, the U.K. code loses the letters F and G and the symbols . and −. The code developed in this paper is very similar to both of these codes but removes these major flaws.

All of the criteria of the Fieldata study are used here. The full spectrum of expansion and contraction among 4-, 5-, 6-, 7- and 8-bit sets is considered in addition. Thus there are the following additional criteria and remarks:

1. A collating sequence has utility in data processing codes containing alphabets; transmission codes do not require such a sequence.
2. A collating sequence has no utility in a 4-bit set.
3. A collating sequence has utility in 5- and 6-bit sets and it is desirable that the sequence correspond to the binary representations.
4. If it is assumed that the 7- and 8-bit sets contain upper and lower case forms of the same alphabet, it is impossible to have the collating sequence match the binary representations, for the case distinction is of lesser significance than the distinction between characters with different meanings. [7, 8]
5. It is not necessary that the full 4-bit set be in 16 contiguous positions in larger sets. It is only necessary that the lowest four bit positions form the dense, unduplicated set. Other bit positions may vary. However, the digits 0-9 (10, 11) should be certainly be grouped contiguously in any set.
6. Punctuation characters have natural delimiting functions and should thus collate low to both the alphabet and digits. These include, but are not limited to:

   blank . , / − : ; ′ ( ) (not in ranked order)
7. Since period and hyphen are natural delimiters, they should be placed low to both alphabets and digits. However, they often serve as radix point and minus sign (which are not delimiters) in the 4-bit numeric set. There must also be a character in this set to serve as a blank; this may or may not print in the 4-bit numeric mode. Therefore any characters of the 4-bit set which are delimiters should be in a different contiguous block than the digits, so they can serve the delimiting function in larger sets. There should be some regular transformation to append bits when expanding to larger sets.
8. All expansion and contraction from and to the various set sizes shall be blind, without knowledge of the meaning of the character assigned to any bit representation, or of contextual adjacency (with the exception of FIGURE/LETTER SHIFT control in going between 5- and 6-bit sets).
9. In all expansion and contraction, MASTER SPACE must remain all 0's and DELETE must remain all 1's. ESCAPE shall always be the second highest code, one less than DELETE; thus all bits except the low order are 1. For paper tape usage, BLANK must be different from MASTER SPACE and therefore shall have all bits 0 except the low order. This guarantees that BLANK, as the primary delimiter, collates low to all other characters. It is also the complement of ESCAPE.
10. All possible caution should be exercised in alphabetic regions to provide maximum expansion for non-English alphabets (> 26 letters).

The 8 bits are represented by B_7 through B_0, high to low order. The 6-bit transmission set will be developed first. Figure 1 shows a modified Fieldata pattern with B_5 not yet assigned, reflecting criterion 9 only. $B_5 = 0$ for Fieldata, $B_5 = 1$ for U.K.

It is now obvious that LS and FS should be opposite ES and DL, not MS and BLANK, in order to maximize the number of punctuation characters following BLANK. Since the decimal digits must have their binary representation equal to the binary value, they must be placed

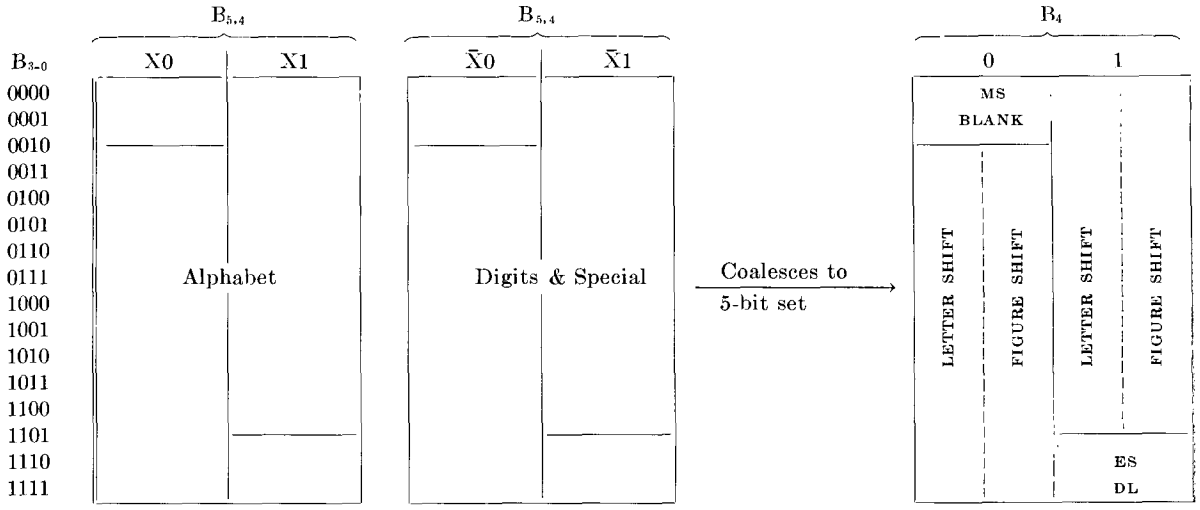


FIG. 1

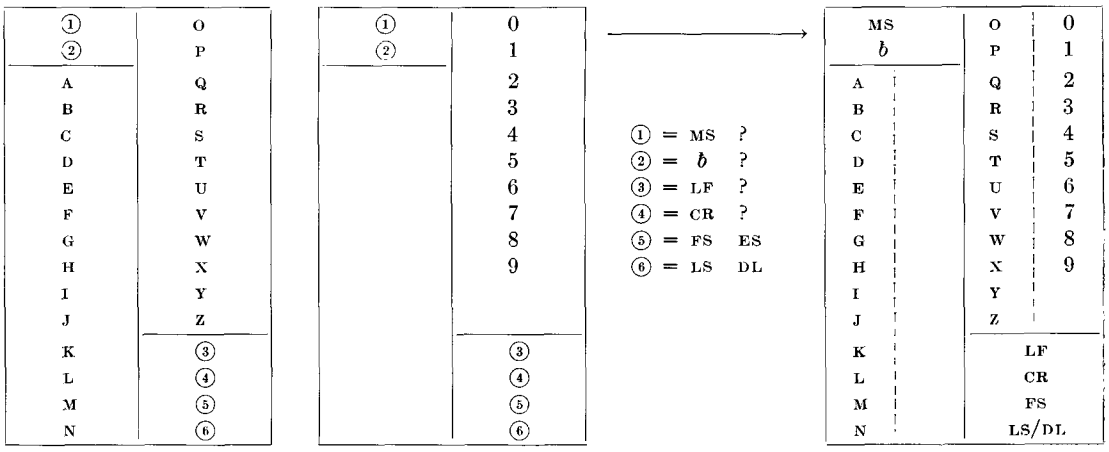


FIG. 2

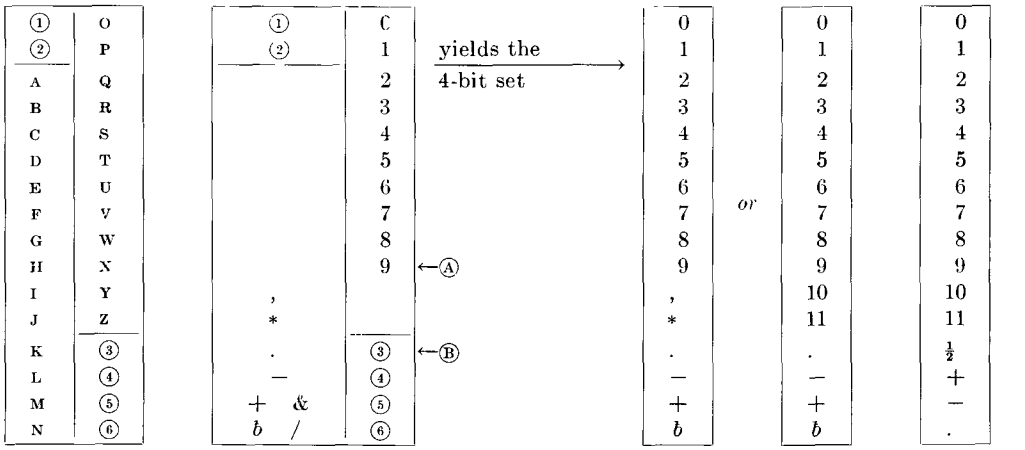


FIG. 3

For	.	–	+ &	/	,	(	)	'	:	?	=	$	⅋ £	;	"	!	*	@	□	%	<	>
Basic CCITT	√	√	√	√	√	√	√	√	√	√	√											
Teletype	√	√	√	√	√	√	√	√	√	√		√	√	√	√	√						
Basic IBM Printers	√	√	√	√	√							√	√				√	√	√	√		
FORTRAN Printers	√	√	√	√	√	√	√	√			√	√					√					
1410, 7070, COBOL	√	√	√	√	√	√	√	√	√		√	√	√	√			√	√	√	√	√	√

FIG. 4

in the $\bar{X}1$ quadrant. This is shown in Figure 2. LINE FEED (LF) and CARRIAGE RETURN (CR) are necessary for character-at-a-time[1] printers associated with existing communication systems. As control signals, they are grouped with other control signals rather than with MS and BLANK, which are essentially informational.

There is space for 20 special characters in the 6-bit set, but four of these must disappear in the 5-bit set.

In conformity to most existing practice, the other six characters of the 4-bit set have been selected as:

. – + (for self-delimiting data fields)

b printing separator (may have graphic representation)

, digit grouping } most expendable for

* indicator for totals, etc. } British pence (10, 11)

The two positions following the digit 9 are not usable for delimiters in the 6-bit set, since they will collate high. These six characters are assigned as shown in Figure 3. The pairs (.,) and (−+) are a distance of two bits apart, for easier error detection. + is used fully interchangeably with &, since & may also take the forms ✝ ⅋ &. / has been chosen as alternate for the BLANK in the 4-bit set, since it can serve very well as a space indicator, as 00/636/505///21.

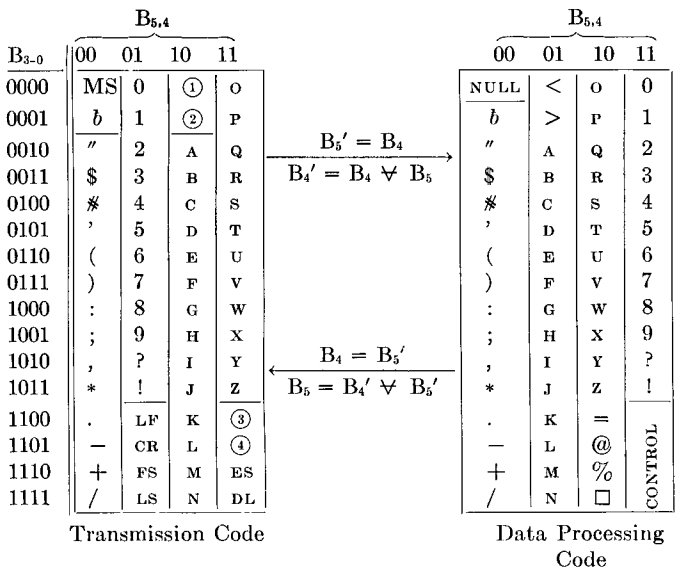


FIG. 5. The Proposed Standard Code

[1] The common term in communications is "page" printer; however, printers which print an entire page at one time must preempt this term.

The basic set is achieved by changing the transform at Ⓐ.

$$B_4 = \bar{B}_3 \vee (\bar{B}_2 \wedge \bar{B}_1)$$

The British set *should* have 10 and 11 immediately following 9. This is achieved by changing the transform at Ⓑ.

$$B_4 = \bar{B}_3 \vee \bar{B}_2$$

Figure 4 gives the special characters specified in existing systems.

A FORTRAN-commercial substitution exists to overcome limited capacity of line printers. The correspondence is:

⅋ to = @ to ' % to (□ to)

The Bell and "who are you" functions are ignored here because they do not warrant individual characters. They are handled best by the ESCAPE mode.

B_5 may now be assigned specifically.

X = 0, $\bar{X}$ = 1 yields a modified FIELDATA which, unless transformed, has punctuation high to the alphabet. This is not logically consistent.

X = 1, $\bar{X}$ = 0 yields modified U.K.

" UNIVAC, MH

" RCA 501

" 704 internal

We will thus choose the latter. This choice also diminishes the number of bits or punches in numeric data, which is most frequent to data processing.

The specific proposal of Figure 5 implies either that:

(a) the data processing code is internal, and the EXCLUSIVE OR mapping takes place at the interface on reading or writing externally on media such as tape or communication lines, which utilize the transmission code, or

(b) the data processing code is merely figurative and represents the effective collating sequence obtained by a simple comparison logic in the machine.

The transmission code folds to the Baudot-like code of Figure 6. It retains all the special characters of present-day Teletype, plus the *. Although ? and !, as effective delimiters, might well precede the alphabet in the data processing code (involving a swap with < and >), to do so would remove ? and ! from the 5-bit transmission code. If the transmission people agree, this change could be considered.

As a 6-bit transmission code, ① to ④ are available. These might be used either for additional control func-

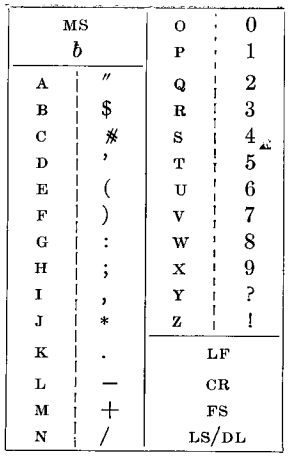
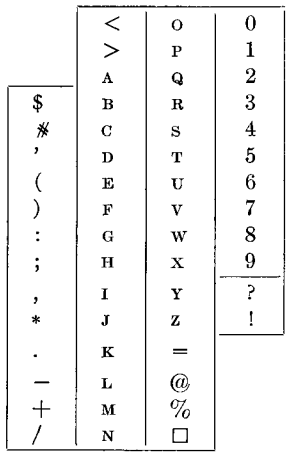
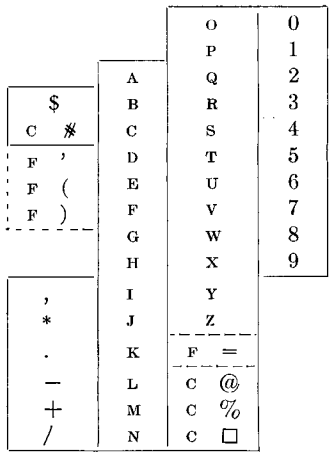


FIG. 6



57 + Blank



48 + Blank

FIG. 7

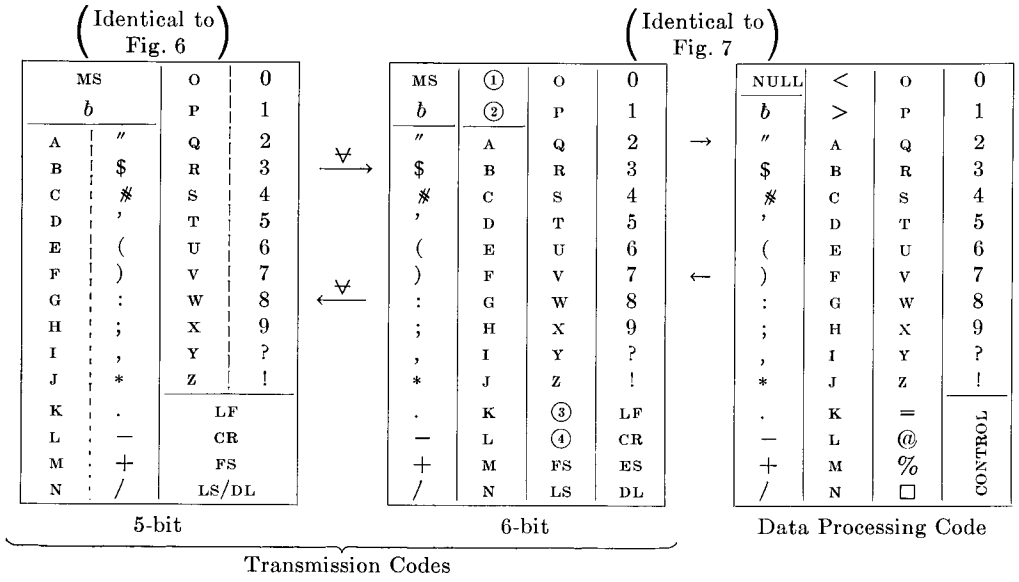


FIG. 8

tions or (preferably) for additional characters of foreign alphabets. FS and LS are also available, once the complete change is made from existing equipment. Fieldata would then use these characters as UPPER CASE (UC) and LOWER CASE (LC) respectively. This provides a representation of a 7-bit code in 6-bit form, just as Baudot represents a 6-bit code in 5-bit form. Therefore the FS-UC and LS-LC correspondences are true, and either mnemonic might be used. Perhaps a new combination would be desirable, as FU (for Figure/Upper) and LL (for Letter/Lower).

= @ % and □ are placed high in the data processing code, and it is assumed they will not be used in control keys. Figure 7 shows the DP code satisfying the last set of special characters of Figure 4, plus the FORTRAN transformation and a 48-character set.

The proposed set has the special characters assigned for reasons other than matching correspondence between the digits and the characters associated with those digits on typewriter keyboards. The reasons are:

1. Some typewriters do not have keys for the digit one (1) or even for the digit (0).

2. There is no such thing as a standard typewriter keyboard in the U.S. There is a proposed British Standard, but the characters placed most uniformly, the left and right parentheses, are above 9 and 0 respectively. This conforms with much practice in the U.S., but 0 must be placed in parallel to MS in accordance with our previous rules.

3. Transmission people sometimes desire the parentheses over the 8 and 9 respectively, but this occurs only in the Luebbert revision of Fieldata (not the original Fieldata) and Ferranti computers.

4. Users normally adjust automatically to any arrangement of special characters after a day's usage.

5. Non-English keyboards differ greatly in this placement. It would be unfair to inflict one of the many English arrangements as an international standard.

6. The FORTRAN-Commercial interchange is accomplished in the proposed set by recognizing

$$B_2 \wedge \overline{B}_4 \wedge (B_3 \equiv B_5),$$

and inverting B_3 and B_5 if this condition is true. Any other arrangement greatly complicates the logical hardware necessary in converting existing printers, probably a more expensive process than converting existing Teletypewriters.

7. Any such correspondence will still require two modes of keyboard logic to generate codes.

If the transmission people could modify existing equipment with an EXCLUSIVE OR function (two relays), a completely common and collatable 6-bit code could exist, as shown in Figure 8, subject to the requirements for expansion to 7- and 8-bit sets.

REFERENCES

1. Electronic Industries Association, Basic Character Set Code, Tentative Standards Proposal 7233, May 1960.
2. LUEBBERT, W. F. Information handling and processing in large communication systems. Tech. Report 099-1, Stanford Electronics Laboratories, Stanford University, 11 July 1960.
3. U. S. Army Signal Corps, Fieldata Equipment Intercommunication Characteristics. 1 August 1959.
4. LUEBBERT, W. F. The design of the new military common-language data code. (Dittoed Copy).
5. British Standards Institution, Committee DPE 149, Draft British Standard for Punched Tape Coding, Part 1, 7 Track Code, AA (DPE) 3543, Sept. 1960.
6. BEMER, R. W. A proposal for character code compatibility. *Comm. ACM 3*, No. 2, Feb. 1960.
7. BEMER, R. W. On the design of extended character sets. 26 Jan. 1961 (Unpublished).
8. SMITH, H. F. JR. AND WILLIAMS, F. A. JR. Remarks on collating sequences. 22 Sept. 1960 (unpublished).

APPENDIX

On the Relative Position of the Alphabet, Numbers and Special Characters in a Code Set Based upon Transmission and Data Processing Characteristics

Consider a 4-quadrant arrangement of 16 states per quadrant in which Q1, Q2, Q3 and Q4 represent the octal codes 00-17, 20-37, 40-57 and 60-77. Consider also four classes of information symbols which may be placed in these quadrants:

- D representing the 10 or 12 digits of the decimal or duodecimal system
- S representing the class of special characters
- A representing a section of the alphabet beginning with the letter A
- Z representing a section of the alphabet ending with the letter Z

If each of these four classes of information is assumed to consist of up to 16 codes, they may be assigned to the 4 quadrants in any of 24 combinations. We now consider these combinations in light of their desirability for data processing and data transmission.

Transmission Considerations

1. Concepts of MASTER SPACE and BLANK are distinct. The term "BLANK" refers to the element of information used to separate words on a printed page. MASTER SPACE occupies the zeroth position.
2. The concept of ERASE or DELETE is represented by the *N*th character.
3. MASTER SPACE, BLANK and DELETE are concepts required in all alphabetic or alphanumeric sets. Further MS and DL always occupy the same relative position in each set.
4. The 64-state code set is to be representable by a 32-state code using a shift mode. In this compressed representation the alphabet is to form one shift and the numbers and special characters the other.

These criteria imply:

01. A and Z cannot fold upon each other (4).
02. Q1 and Q4 cannot fold upon each other (1, 2, 3).
03. MASTER SPACE must be in Q1. (1)

Data Processing Considerations

5. The digits are represented by their pure, natural binary equivalents. Since only four bit positions are necessary to represent up to 16 states, any additional bit position in a given set must contain the same pattern of 1's and 0's for each digit.
6. No symbol other than MASTER SPACE ranks lower than BLANK in the collating sequence.
7. The alphabet is dense in collation.
8. Certain field-separating symbols including BLANK must rank lower than the alphabet in collating.

These criteria imply:

04. D cannot be in the quadrant which contains or folds on MASTER SPACE. Otherwise 0 and MASTER SPACE would become identical in some code set. (5)
05. Z cannot occupy Q1. (7)
06. MASTER SPACE and BLANK must appear as adjacent characters in the same quadrant. Otherwise some symbols will be less than BLANK or MS will have a rank higher than BLANK. (6)
07. BLANK cannot be associated with A. Otherwise some special symbols would either be lower in collating than BLANK, or field-breaking symbols would be higher in collating than the alphabet which they are intended to separate. (6, 7)
08. (7) and (8) immediately above imply that A cannot occupy Q1 since MS must be located here.

Application of Rules to the Combinations of D, A, Z, and S on Q1, Q2, Q3, and Q4

The requirements 04, 08 and 05 remove from consideration the 18 combinations beginning with D, A and Z respectively. In addition any combination in which A does not precede Z can result in a non-dense alphabet.

(*please turn to page 225*)

TABLE 1

Deviates of the Normal Function in Octal Corresponding to the Cumulative Area From .5 to 1.0

Area (Octal) Scaled 2^8	Normal Deviate (Octal) Scaled 2^{28}	Area (Octal) Scaled 2^8	Normal Deviate (Octal) Scaled 2^{28}
200	00 00000 00000.	300	00 12625 32704.
	00 00120 15457.		00 12772 42473.
	00 00240 33636.		00 13140 41660.
	00 00360 53002.		00 13307 32316.
	00 00500 73431.		00 13457 16144.
	00 00621 16043.		00 13627 77435.
	00 00741 43141.		00 14001 60730.
	00 01061 73214.		00 14154 44377.
210	00 01202 26537.	310	00 14330 34471.
	00 01322 66042.		00 14505 34031.
	00 01443 32032.		00 14663 45573.
	00 01564 03006.		00 15042 74314.
	00 01704 61253.		00 15223 42707.
	00 02025 45515.		00 15405 34463.
	00 02146 40500.		00 15570 55071.
	00 02267 42513.		00 15755 27637.
220	00 02410 54064.	320	00 16143 40227.
	00 02531 75526.		00 16333 12607.
	00 02653 30005.		00 16524 33524.
	00 02774 73427.		00 16717 27240.
	00 03116 50535.		00 17114 02241.
	00 03240 40077.		00 17312 41757.
	00 03362 42443.		00 17512 73747.
	00 03504 60346.		00 17715 25671.
230	00 03627 12161.	330	00 20121 65564.
	00 03751 60466.		00 20330 42335.
	00 04074 44256.		00 20541 43267.
	00 04217 45711.		00 20755 00012.
	00 04342 65610.		00 21173 00365.
	00 04466 24554.		00 21413 55747.
	00 04612 03410.		00 21637 22316.
	00 04736 02547.		00 22065 70106.
240	00 05062 22626.	340	00 22317 52300.
	00 05206 64474.		00 22554 65705.
	00 05333 51003.		00 23015 50463.
	00 05460 60415.		00 23262 21226.
	00 05606 13615.		00 23533 00141.
	00 05733 73512.		00 24010 07632.
	00 06062 01012.		00 24271 74500.
	00 06210 34430.		00 24560 64671.
250	00 06337 16705.	350	00 25055 11231.
	00 06466 30775.		00 25357 26315.
	00 06615 73671.		00 25667 74207.
	00 06745 70150.		00 26207 36451.
	00 07076 16612.		00 26536 46016.
	00 07227 00465.		00 27076 02014.
	00 07360 16636.		00 27446 51276.
	00 07511 72144.		00 30031 33024.
260	00 07644 03467.	360	00 30427 40773.
	00 07776 54160.		00 31042 23633.
	00 10131 65174.		00 31473 35502.
	00 10265 37466.		00 32144 76235.
	00 10421 54234.		00 32641 23053.
	00 10556 34507.		00 33363 41636.
	00 10713 61765.		00 34137 46520.
	00 11051 55313.		00 34752 51075.
270	00 11210 20025.	370	00 35633 37754.
	00 11347 33317.		00 36573 54572.
	00 11507 21011.		00 37631 00460.
	00 11647 62131.		00 41010 33357.
	00 12011 00143.		00 42354 42517.
	00 12152 74634.		00 44204 72526.
	00 12315 51606.		00 46534 51133.
	00 12461 10461.		00 52437 23555.

STANDARDS—Continued from page 217:

As can be seen, this is not the only method of applying the stated rules. Despite the order taken the rules reduce to three the number of acceptable combinations for data processing and transmission. These combinations are S, A, Z, D which is followed by IBM and S, D, A, Z which is advocated in the United Kingdom. Also possible is S, A, D, Z.

The FIELDATA arrangement A, Z, S, D is not acceptable. First, BLANK is associated with A, which means the delimiting special characters will collate higher than the alphabet. Second, if BLANK is not associated with MASTER SPACE in Q1 but is in the second position of Q3, one symbol (the character in the first position of Q3) other than MASTER SPACE must have a rank less than BLANK.

This examination of the possible combinations of S, D, A, Z merely indicates which arrangements should be further investigated with view of their expansion characteristics in sets of more than 6 bits. The analysis is intended to remove much of the confusion which has existed as to what combinations are possible and desirable for expansion and contraction. Thus, the arrangements S, A, Z, D and S, D, A, Z will be given further analysis and a choice between them made on the facility of their expansion and contraction characteristics.

Folding Considerations

At this point the methods of folding must be considered. The four quadrants may be folded in one of the two ways by the removal of either B5 or B4.

Q1 on Q2 and Q3 on Q4
Q1 on Q3 and Q2 on Q4

When the folding consideration is applied to the remaining combinations of S, A, D, Z we find,

	Q1 on Q2	Q1 on Q3
SDAZ	Note 1	Possible
SAZD	Note 2	Note 4
SADZ	Note 3	Note 1

NOTE 1. The digits cannot fold on MASTER SPACE.

NOTE 2. The transformation characteristic between 5 and 6 bits is dependent upon the combination being treated as well as the shift.

NOTE 3. This is possible by treating B_4 of the 6-bit representation as the mode bit. However, this leads to a non-dense alphabet.

NOTE 4. Z cannot fold on MASTER SPACE.

Conclusion

From the consideration of data transmission and data processing criteria we are led to a code organization of S, D, A, Z. This organization, however, should not be considered as giving the collating sequence.