"METHODS OF INFO. HANDLING"
2ND EDITION DRAFT

NEW TEXT

DRAFT

1975?

3. Coding: The Indexing Shorthand

A. The Statistical Nature of English Words

1. Initial Letter Usage -- Continuous Text
2. Initial Letter Usage -- Subject Headings
3. Initial Letter Usage -- Surnames
4. Terminal Letter Usage -- Continuous Text
5. Composite Letter Usage
6. Letter Usage by Position
7. Bigram (Letter Pair) Usage
8. Distribution of Subject Heading Lengths and Surname Lengths
9. Statistical Summary

B. Derived Coding

1. Derived Coding Methods
2. Evaluation of Derived Coding Methods

C. Assigned Codes

1. Text and Communication Codes
2. Journal Title Codes
3. Book Identification Codes
4. Subject Classification Codes
5. Chemical Codes
6. Geographic Place Name Codes
7. Organization Codes
8. Account Identification Codes
9. Biographic File Codes

D. Equal Internal Coding

E. Special Coding Methods

1. Check Digits

2. Prime number Coding

3. Superimposed Coding

References

The Statistical Nature of English Words
~~Methods for Systematically Abbreviating English Words and Names~~
~~Prime Number~~ Derived Coding
~~Superimposed~~ Assigned Coding
Equal Interval Coding
Special Coding Methods

3

Coding: The indexing shorthand

In addition to establishing the over-all organization of the file collection, the system designer must decide how the classification or indexing information is to be represented, whether on paper or in some machine-readable form. In most cases, coding is a shorthand representation of the indexing information, and can help system operation by reducing clerical effort and storage requirements, and by streamlining the search and retrieval process. The choice of coding method becomes critical if automatic or semi-automatic processing or handling of the file or indexing information is to be used, and this special topic of machine language representation is examined in Chapt. 4. There are, however, several special and useful coding topics that have been described previously only in widely scattered sources. These less-well-known coding techniques and background information have been collected into this chapter. Other coding topics are discussed in Chapts. 5, 6, and 7 as they relate to specific types of equipment or media. Further information is available from the sources listed at the end of the chapter.

A. THE STATISTICAL NATURE OF ENGLISH WORDS

A. The Statistical Nature of English Words

Frequency distributions for the alphabetic letters in English text material and in special collections of words (e.g., name lists) have been determined by many investigators.[1] A summary of many of the results of these studies is presented here.

1. Initial Letter Usage -- Continuous Text

Table 3-1 gives three different rankings of frequency of occurrence of initial letters in continuous English text material. These three studies show relatively close correlation, and agree that T, A, and O are the three most frequent characters for initial letters.

2. Initial Letter Usage -- Subject Headings

Two studies of subject words also show relatively close correlation, and agree as shown in Table 3-2 that S, P, and C are the three most frequent characters for initial letters in subject headings.[2] A less complete study of the initial letter frequency of four major scientific and technical thesauri provided additional data that tended to confirm the rankings and distributions given in Table 3-2.[3] (3. Papier, L.S., and T.L. Lin, "First Letter Frequency of Related Term References in Four Technical Thesauri," Proceedings of the American Documentation Institute. 1966 Annual Meeting. Santa Monica, California, Adrianne Press, pp. 409-417.) Taken as a group, the rankings for subject words are different from those for continuous text. This difference would be expected, primarily because of the large number of prepositions and other

[1] Pratt, F., *Secret and Urgent, The Story of Codes and Ciphers* (Blue Ribbon Books, Garden City, New York, 1942).

Ohaver, M. E., *Cryptogram Solving* (Stoneman Press, Columbus, Ohio, 1933).

Cox, G. J., et al., "Recent Developments in Keysort Cards," *Journal of Chemical Education*, Vol. 24, pp. 65-70 (February 1947).

Smith, L. D., *Cryptography—The Science of Secret Writing* (W. W. Norton Co., New York, 1943).

Gaines, H. F., *Cryptanalysis* (Dover Publications, New York, 1956).

Griffith, R. T., "The Minimotion Typewriter Keyboard," *Franklin Institute Journal*, Vol. 248, No. 5, pp. 399-436 (November 1949).

[2] Bourne, C. P., and D. F. Ford, "A Study of the Statistics of Letters in English Words," *Information and Control*, Vol. 4, No. 1, pp. 48-67 (March 1961).

Ohlman, H., "Subject Word Letter Frequencies with Applications to Superimposed Coding," *Proceedings of the International Conference on Scientific Information*, Vol. 2, pp. 903-916 (National Academy of Sciences, Washington, D.C., 1959).

TABLE 3-1

Comparative Ranking by Frequency of Occurrence of Initial Letters in English ~~Words~~ Text

~~Continuous Text~~ (Pratt)	~~Continuous Text~~ (Ohaver)	~~Continuous Text~~ (Smith)
T	T	T
A	A	O
O	O	A
S	S	W
W	H	B
I	I	C
H	W	D
C	C	S
B	B	F
F	P	M
P	F	R
M	D	H
R	M	I
E	R	Y
L	E	E
N	Y	G
D	N	L
U	L	N
G	U	P
Y	G	U
J	V	J
V	K	K
Q	J	
K	Q	
X	Z	
Z	X	

TABLE 3-1

Comparative Ranking by Frequency of Occurrence of Initial Letters in English ~~Words~~ Subject Headings

~~Subject Words~~ (Bourne-Ford)	~~Subject Words~~ (Ohlman ~~p. 92~~)
S	C
P	S
C	P
T	A
A	B
D	M
M	T
R	R
B	E
I	F
E	D
F	G
H	H
L	I
G	L
W	N
N	W
O	O
V	V
J	U
U	J
K	K
Q	Q
Y	Y
Z	Z
X	X

small words (e.g., a, and, in, is, the, to, with) that appear in continuous text.

3. Initial Letter Usage -- Surnames

Table 3-3

Table 3-3 shows that five different studies of initial letters in people's names were in relatively close agreement.[4] (4. Department of Health, Education , and Welfare. Social Security Administration. Bureau of Old Ages and Survivors Insurance. Division of Accounting Operations. Preliminary Report of Distribution of Surnames in the Social Security Account Number File. July 1, 1957. Report No. 6288.) All the studies showed that S, M, B, H, and C are the most frequent characters for initial letters.

Fn 4

4. Terminal Letter Usage -- Continuous Text

The three studies of continuous text agreed that E, S, D, T, and N were the most frequent characters for terminal letters, and were in relatively close agreement for the remainder of the characters. The rankings of terminal letters in the study of subject words were distinctly, but not radically, different from those for continuous text. All the studies agreed that E was the most common terminal letter.

5. Composite Letter Usage

Table 3-4

Table 3-4 lists data from six different studies of the composite rankings of letter usage. Composite rankings are based on the relative frequency of occurrence of letters in collections of words or in text and do not consider the position of the letters within the words. The four

TABLE 3-3

Comparative Ranking by Frequency of Occurrence of Initial Letters in ~~Proper Names~~ Surnames

Bourne-Ford [a]	Bourne-Ford [b]	Ohlman, ~~p. 89~~	Cox *et al.*, ~~p. 88~~	Social Security Administration [c]
B	S	S	S	S
S	B	B	B	M
M	M	M	M	B
H	H	H	H	H
C	C	C	C	C
W	W	D	W	W
R	R	G	R	R
L	G	K	P	G
P	L	L	L	P
G	P	R	G	D
D	D	P	D	L
K	F	W	F	K
F	K	A	K	F
T	T	F	T	T
A	A	T	A	A
J	J	E	E	J
E	E	N	J	E
N	N	V	N	N
O	O	J	V	O
V	V	O	O	V
Y	I	I	Y	Y
I	Y	Z	I	Z
Z	Z	U	Z	I
U	U	Y	U	U
Q	Q	Q	Q	Q
X	X	X	X	X

[a] Based on a detailed study of 8207 names.

[b] Based on an independent, but less-detailed study of approximately 63,000 names.

[c] Based on study of over 117 million names.

TABLE 3-8

Comparative Ranking by Composite * Frequency of Occurrence of Letters in English Words and ~~Proper~~ Surnames

Subject ~~Heading Words~~ (Bourne-Ford)	Continuous Text (Pratt, p. 252)	Continuous Text (Ohaver, p. 25)	Continuous Text (Griffith, p. 426)	Continuous Text (Smith, p. 153)	~~Proper~~ Surnames (Bourne-Ford)
E	E	E	E	E	E
I	T	T	T	T	A
R	A	A	A	O	R
O	O	O	O	A	N
A	N	N	N	N	L
T	R	I	I	I	O
N	I	S	S	R	I
S	S	R	R	S	S
L	H	H	H	H	T
C	D	L	L	D	H
P	L	D	D	L	D
M	F	C	C	C	M
D	C	U	U	W	C
U	M	P	M	U	B
H	U	F	F	M	G
G	G	M	Y	F	U
Y	Y	W	W	Y	W
B	P	Y	G	G	Y
F	W	B	P	P	J
V	B	G	B	B	K
K	V	V	V	V	P
W	K	K	K	K	F
X	X	Q	X	X	V
Z	J	X	J	Q	Z
J	Q	J	Q	J	X
Q	Z	Z	Z	Z	Q

* These are all composite samples which did not consider the position of the letters within the words.

studies of continuous text material showed remarkably close agreement, and all showed that E, T, A, O, and N were the most frequent letters. The rankings for subject words and for surnames differed from each other and from the rankings for continuous text material. The only point of complete agreement in all six studies was that E was the most common letter. A more graphic comparison of the composite distributions for subject words and for surnames is given in Fig. 3-1, which shows the actual frequencies of usage, as well as the rankings. Coding systems have been developed to provide very compact representations of alphabetic characters by taking advantage of character frequency data such as that given in Table 3-3. (Ramamoorthy, C.K., "Document Compaction by Variable Length Encoding," Proceedings of the American Documentation Institute. 1964 Annual Meeting. Vol. 1, Spartan. pp. 507-513.)

6. Letter Usage by Position

Data is available to describe the rankings and frequency distributions for the characters of subject words in a particular letter position (i.e., the frequency distribution for the initial letter of the word, the second letter of the word, etc.). (Bourne and Ford cited earlier.) Some interesting patterns are noted. For example, the distribution of the first, second, and possibly the third, letter positions have relatively unique patterns, and those of the remaining letter posistions have generally the same pattern. In the second letter position, the vowels are the most common characters. The ranking is E, O, A, I, R, U. Excluding the R, the group of five vowels accounts

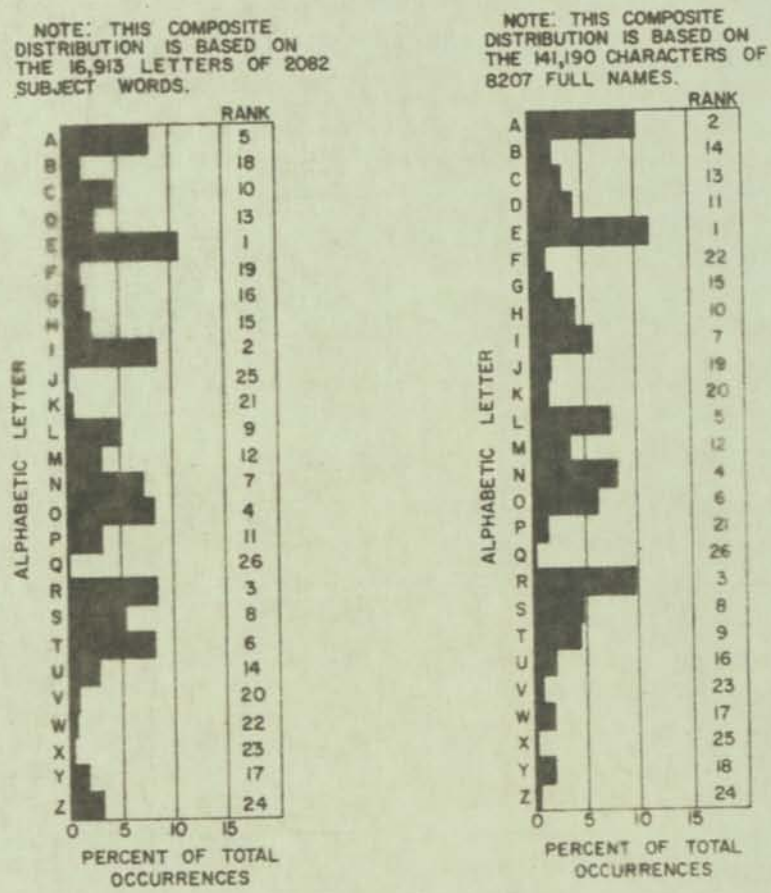


FIG. 3-1 Composite distributions and rankings for the alphabetic letters in subject headings ~~words~~ and ~~people's names~~ surnames.

for over 61 percent of the letter usage in that position. The same type of data is available to describe the rankings and frequency distributions for the characters of surnames within a particular letter position. As in the case of the subject words, the distributions of the first and second letter positions have relatively unique patterns, while those for the remainder of the letter positions have much the same pattern. The vowels dominate the second letter position of surnames, as they did with subject words. In this case, the vowels (ranked A, O, E, I, U), account for over 70 percent of the character usage in the second letter position. There is general agreement between the patterns for corresponding letter positions of subject words and surnames.

7. Bigram (Letter Pair) Usage

Table 3-5 lists data from seven different studies of rankings of bigram occurrences. The five different studies of continuous text and telegraph text agreed in many respects (e.g., that TH was the most common bigram). The bigram rankings for subject words and for surnames were different (except in agreeing that ER was the most common bigram) and somewhat different from any of the other five rankings. However, because of the large number of bigram possibilities, it is more difficult to make a quantitative judgment of the degree of similarity between bigram rankings than it is for the ranking of single alphabetic characters. The frequency of occurrence of various bigrams in subject headings and surnames, and their corresponding rankings are reported in the literature. (Bourne & Ford, IFCN, article cited earlier in this chapter, fn.2.)

TABLE 3-5

Comparative Ranking by Composite [a] Frequency of Occurrence of the Most Common [b] Bigrams in English Words and Proper Names

Continuous Text (Pratt, p. 260)	Military Telegraph (Pratt, p. 260)	Continuous Text (Gaines, p. 218)	Continuous Text (Smith, p. 153)	Continuous Text (Ohaver, p. 26)	Subject Words (Bourne-Ford)	Proper Names (Bourne-Ford)
TH	TH	TH	TH	TH	ER	ER
HE	AN	HE	ER	HE	ON	AR
AN	RE	AN	ON	ER	TI	AN
RE	HE	IN	AN	IN	IN	ON
ER	ON	ER	RE	AN	TE	LE
IN	IN	RE	HE	ON	AT	EN
ON	ER	ES	IN	RE	RA	LL
AT	AT	ON	ED	AT	RO	HA
ND	TO	EA	ND	ED	IO	EL
ST	OU	TI	HA	ST	AN	NE
ES	OR	AT	AT	ND	OR	RO
EN	ND	ST	EN	ES	IC	AL
OF	TA	EN	ES	HA	RE	IN
TE	EN	ND	OF	OF	TR	MA
ED	ES	OR	OR	EN	AL	RI
OR	NT	TO	NT	OR	EN	CH
TI	TI	NT	EA	AS	AR	LI
HI	HA	ED	TI	NT	NT	RE
AS	ST	IS	TO	TI	LE	TH
TO	RO	AR	IT	TO	ET	BE
AR	AR	OU	ST	IT	NE	AM
OU	CO	TE	IO	EA	RI	ES
IS	TE	OF	LE	NG	ST	IL
IT	IT	IT	IS	OU	IT	RT
LE	ME	HA	ON	AR	ES	RD
NT	TT	SE	AR	ET	CO	RA
RI	OM	ET	AS	HI	LA	LA
SE	VE	AL	DE	TE	DE	OR
HA	UR	RI	RT	RO	ME	IC
AL	ED	NG	VE	NE	LI	TE

[a] These rankings did not consider the position of the letter pairs within a word.

[b] The bigrams which consisted of a letter and a space were not considered for these rankings.

8. Distribution of Subject Heading Lengths and Surname Lengths

Sample distributions of subject heading lengths and of surname word lengths are illustrated in Figs. 3-2 and 3-3, respectively. The subject words represented single-word descriptors (e.g., "magnetic," "optical") which were used to index a collection of technical documents in a coordinate indexing system. The surnames consisted of the last name, followed by some combination of the first and middle names and prefixes. In single word descriptors, one study showed that over 95 percent of the vocabulary could be represented by a word length of 13 characters. For surnames, one study showed that approximately 95 percent of the names could be represented by a string of 25 characters. One other study reported that in a study of 3½ million surnames, none were found that were larger than 15 characters.

9. Statistical Summary

In general, the statistical characteristics of usage of the alphabetic letters are markedly different for three different categories of English words -- continuous text material, surnames, and subject headings. Several independent studies of continuous text material show close agreement for a number of characteristics, suggesting that the reported data are representative of the general nature of continuous English text material.

In subject words (two studies), S, P, and C were the three most frequent initial letters. In continuous text (three studies), T, A, and O were the three most frequent initial letters. In surnames (five studies), S, M, B, H, and C were the most frequent initial letters.

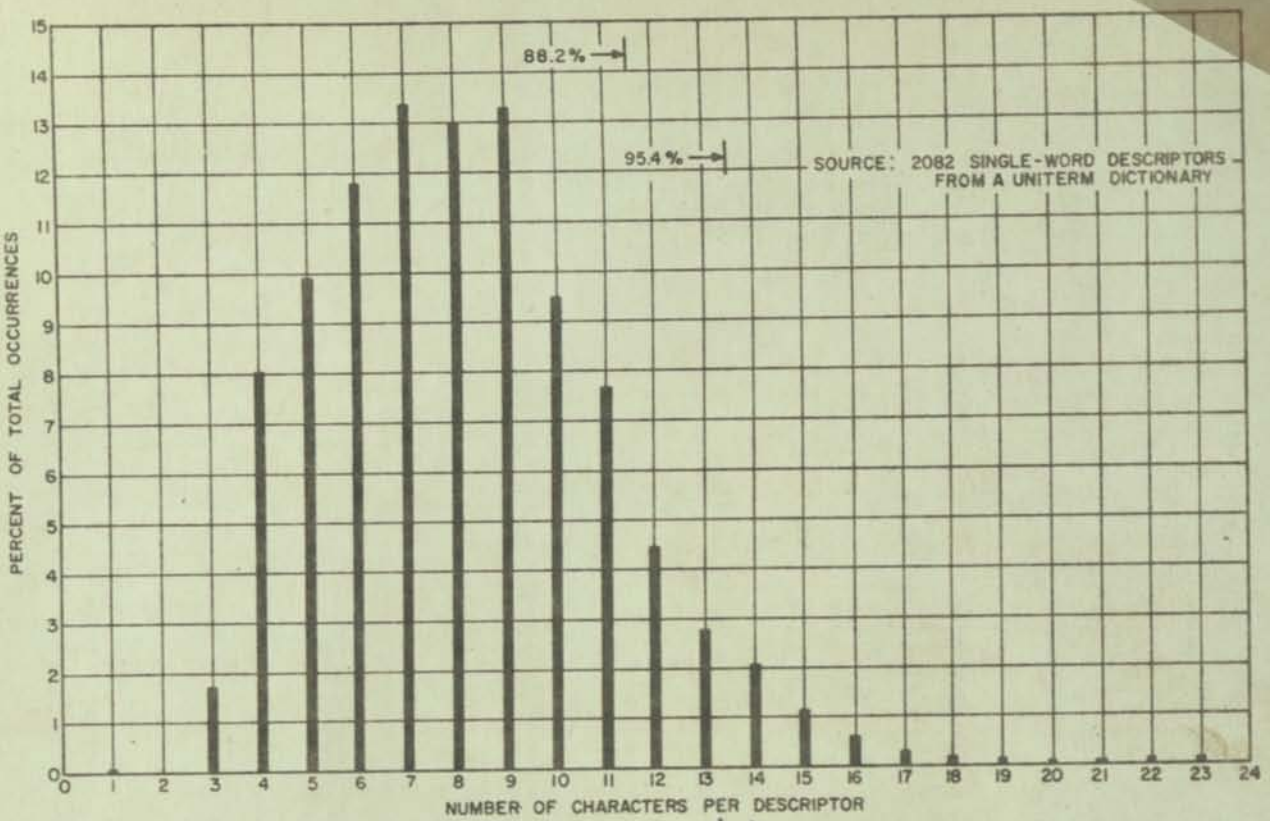


FIG. 3~~A~~2 Distribution of subject ~~word~~ heading lengths.

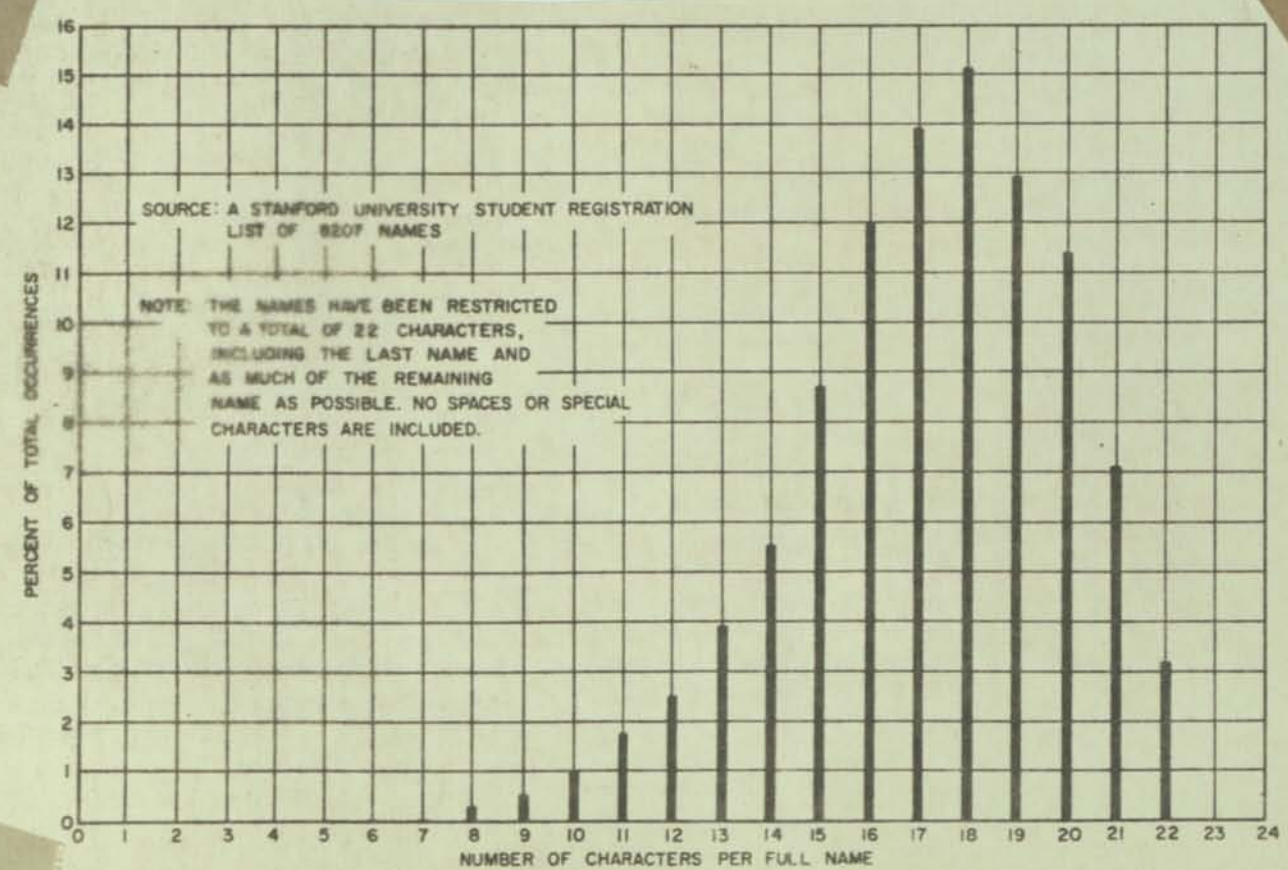


FIG. 3~~-B~~3 Distribution of ~~proper name~~ surname lengths.

In subject headings (one study) E, N, R, and S were the most frequent terminal letters. In continuous text (three studies), E, S, T, and D were the most frequent ~~composite~~ terminal letters. In subject headings (one study), E, I, R, and O were the most frequent composite letters. In continuous text (four studies), E, T, A, O, and N were the most frequent composite letters. In surnames (one study), E, A, R, and N were the most frequent composite letters.

In subject headings and surnames (one study each), the first, second, and possibly the third letter positions had relatively unique letter distribution patterns, and the remaining letter positions had patterns that closely resembled each other.

In subject headings (one study), the most frequent bigrams were ER, ON, TI, IN. In continuous text and telegraph text (five studies), TH, AN, HE, and RE were some of the most frequent bigrams. In surnames (one study), the most frequent bigrams were ER, AR, AN, and ON.

B. Derived Coding

There are many instances in which it is useful to abbreviate[6] (fn.[6] This discussion refers to the "abbreviation" of words, but actually it is concerned with "coding" the words, or transforming the source words into new patterns. It happens that in satisfying some of the objectives of code design the word is abbreviated. However, some of the schemes and transformations described in the following sections bear little resemblance to what might normally be considered as abbreviation.) English words such as surnames, place names, street addresses, proper nouns, or continuous text material. Common subjects for abbreviation are addresses in telephone directories, customer account number identifications in some data processing systems, and commercial teletype or ham radio vocabularies. Abbreviation or coding schemes provide more efficient use of the storage media in computers, punched card systems, and other storage and processing systems. They may also improve the efficiency of communication systems by decreasing the number of characters transmitted for the same basic message. If the abbreviations are to be made automatically or semiautomatically, they must be obtained by some systematic method. Many methods have been suggested for abbreviating words, and a few of them have been empirically tested with representative data to obtain meaningful performance comparisons. The objective of most of these studies has been to develop methods for the systematic coding of words which satisfy the following requirements:

1. Each word should be coded to require as little storage space as possible.

2. Each word should retain the same degree of discrimination and uniqueness that it had in the original sample. (For example, if there were 2082 unique words in the original list of words or names, hopefully there will still be 2082 uniquely identifiable items after abbreviation.)

3. Each word should retain some mnemonic similarity to the original word, where possible.

4. It should be possible to systematically transform each abbreviated word back into the original word when desired.

5. It should be possible to use the abbreviated words as a basis for sorting the original words into nearly correct alaphabetic order.

These are idealized requirements which have not been met in full by any of the methods used to date.

The abbreviation procedure may or may not rely on prior detailed knowledge of the body of words to be abbreviated. A method based on prior detailed knowledge may be called an assigned code system, and a method not requiring such knowledge is called a derived code system. From the standpoint of uniqueness and storage space, one of the most efficient schemes would be an assigned code method that used a previously established table of code numbers or abbreviations for each word that

would subsequently be encountered. This is the most common coding approach. The scheme would require the initial development of a table and continued reference to it. In contrast, derived code systems would generally accept and abbreviate or code any word presented to the system, without a table lookup or the use of detailed tables generated in advance.

As illustrated in Table 3-6, the 26 characters of the English alphabet offer a large number of coding possibilities, so that a few alphabetic characters are capable of providing a large degree of discrimination. Even if the non-pronounceable letter combinations were removed, there would still be a large number of code combinations. The use of a few alphabetic characters in an automobile licence number, for example, provides a large number of possible codes with just a few letters. A compilation of 1805 three-letter words and 19,005 four-letter words has been published which have the special property of being self-demarcating.[7] (Luhn, H. P., Self-Demarcating Code Words, 2nd ed. (IBM Corp., New York, 1953).) That is, the words are serially unique, disjunctive in combination, and reasonably pronounceable. A computer program has even been written to generate pronounceable three-letter and four-letter words.[8] (Leiner, A. L., and W. W. Youden, "A System for Generating Pronounceable Names Using a Computer," JACM, Vol. 8, No. 1, pp. 97-103 (January 1961).)

Another version of automatic or derived coding is the computation of computer storage addresses from the data itself, as a means for

6

TABLE 3-A

Number of Unique Codes Possible with a Series of Alphabetic Letters

Number of Characters Used in the Abbreviation	Number of Unique Codes Possible
1	26
2	676
3	17,576
4	456,976
5	11,821,376
n	26^n

speeding up the storage and retrieval of a specific item in a large list or table. Examples of some of the transform methods used for this address computation have been described in the literature.[9] (Gall, R. G. and D. E. Brotherton, "Associative List Selector." Chapt. 5 report RADC-TR-66-281 of the Goodyear Aerospace Corp., Akron, Ohio, AD-802993. Oct. 1966.)

1. Derived Coding Methods

As mentioned earlier, the derived coding methods generate a code directly from the source material without the necessity for table look-ups.[10] (In this text, a "table look-up" is defined as the operation of looking up some item in a table or index, given a particular search key. For example, to look for a specific telephone number in a telephone directory, given a person's name, would be considered a table look-up operation.) In nearly all the methods described in this section, the initial letter is retained, and the scheme is not executed if the word length is initially less than or equal to the desired length of the code word. Each of 13 different methods is briefly described below.

a) Simple Truncation of the Right End. Starting from the right end of the word, drop off letters until the

required word size is obtained. No exceptions or other operations are allowed. This very simple operation is often used to derive some library call numbers from author names.

b) *Simple Truncation of the Left End.* This process is the same as the simple truncation of the right end, except that the letters are dropped from the left end of the word.

c) *Elimination of Vowels.* Starting from the right end of the word eliminate vowels (a, e, i, o, u) until the desired word length is reached. If the word cannot be sufficiently shortened by this method, use simple truncation of the right end to reach the desired length after the vowels are eliminated. The common method of manual note taking referred to as "speedwriting" uses this type of coding (plus the elimination of other characters) in order to reduce the amount of material to be written.

d) *Selective Drop-Out by Letter Position.* Starting from the left end of the word, eliminate every *n*th letter. (Each different *n* constitutes a different technique and results in different efficiencies.) The use of this technique to reduce a word to three characters by dropping out every second character (i.e., $n = 2$) is shown in the two examples below.

First pass	ABLATION	ACCEPTANCE
Second pass	ALTO	CPAC
Final result	ATO	APC

e) *Selective Drop-Out by a Single Ranking of Letter Usage.* Using an empirical ranking of the composite[11] frequency of usage of the 26 alphabetic characters, such as the ones given in Table 3-4, eliminate the most common letters until the desired word length is reached.[12] In case two letters are equally common, remove the right-hand letter first. Because of the different source material used in previous empirical studies, the published frequency distributions may not accurately represent the statistical nature of the new data to be abbreviated or processed. It may be necessary to test or determine the frequency distribution of the new input data before using any of the abbreviation schemes that require a frequency distribution.

f) *Selective Drop-Out by Separate Rankings of Character Usage for Each Letter Position.* Using a separate empirical ranking of letter usage for *each* letter position, examine each word to determine the respective ranking of each of the characters present. Then remove the letters in the order of their frequency until the desired word length is obtained. In case two letters have equal rank, remove the right-hand letter first.

It has been found that there is a marked similarity in the ranking and distribution of most of the characters, regardless of their letter position ~~(see Figs. 3-2 and 3-3)~~. However, for both subject words and proper names the distributions of the first two, and possibly three, letter positions are each markedly different from the distributions of all the other letter positions. In particular, the frequency of vowels in the second letter position suggests that this position does not provide a great deal of discrimination.[13]

g) *Selective Drop-Out by a Single Ranking of Bigram (Letter Pair) Usage.* Using an empirical ranking of the composite frequency of occurrence of bigrams (i.e., ignoring the position of bigram occurrence within the words), examine each word to determine the ranking of all the possible letter pairs (adjacent letters only). The elimination process considers the association of each letter with the neighboring letters and rejects the letters which make up the most common bigrams. With the exception of the initial letter (which is retained ~~by policy~~) each letter contributes to two different bigrams. The total ranking of the frequencies of these two bigrams is the selection criterion for that particular letter. Each letter is evaluated in this way to derive its index or measure of discriminative power. Those with the least discriminative power are removed to

[11] That is, an average frequency of usage regardless of the letter positions. There are distinct differences in the character frequency distributions for the various letter positions (Figs. 3-2 and 3-3).

[12] Luhn, H. P., "Superimposed Coding with the Aid of Randomizing Squares for Use in Mechanical Information Searching Systems," Chapt. 23 in *Punched Cards—Their Application to Science and Industry*, 2nd ed., Casey et al. editors (Reinhold Publishing Corp., New York, 1958).

[13] This has been recognized by other workers, notably Cox, Casey, and Bailey, who have designed punched card codes (for coding proper names) that either ignored the second letter or provided separate coding spaces for each vowel and one or two additional code spaces to take care of the remaining 21 characters (Cox, G. J., et al., cited previously fn. 1).

arrive at the desired word length. The process is shown by the following example, reducing the word ACETYLENE to four characters:

Bigram rankings		66	58	25	108	164	24	20	26	2	
Letter	A	C	E	T	Y	L	E	N	E		(space)
Letter index		124	83	133	272	188	44	46	28		
Final result	A	T	Y	L							

It would also be possible to operate with trigrams or larger groups of letters, and consideration could be given to letter pairs that are separated by one or more letters. It would also be interesting to examine the use of a separate bigram ranking for each letter position.

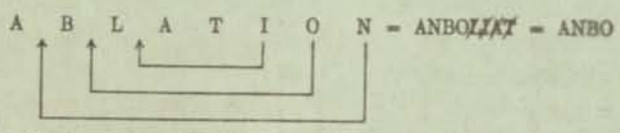
Truncation after Elimination of the Second Character. Eliminate the second character of the word, counting from the left side of the word. Then perform simple truncation, starting at the right end of the word. No exceptions or other operations are allowed. This modification of the simple truncation scheme should benefit by omission of the second letter with its weak discriminative power.

The Use of a Check Digit (or Letter). The discriminative power of any of the preceding coding schemes can probably be enhanced by using a check digit or check letter to describe the letters that have been eliminated.[14] For example, an abbreviation technique may reduce COMPUTERS and COMPUTATION to the same root, COMPUT, with a resulting loss in discrimination. Using a check letter, the original words would abbreviate to COMPU *, the letter in the asterisk position generally being different for each original word. There are several ways to generate a check letter; (1) add the numbers that represent the alphabetic position of the eliminated letters (e.g., A = 1, B = 2, . . . , Z = 26); (2) add the numbers that represent the frequency rankings of the eliminated letters (e.g., E = 1, A = 2, . . . , Q = 26 for names); and (3) count the number of eliminated letters. If the number resulting from any of these operations is larger than 26, then cast out 26 as many times as possible; the remainder (less than 26) is then used as the entry in a table of check letters. The table of check letters could be a simple listing of the alphabet. One method for deriving a check digit is shown in the following example, in which the word ABLATION is reduced to a total of four characters by simple truncation and the addition of a check letter:

									1+20+9+15+14=59=26+26+7=H
Original word	A	B	L	A	T	I	O	N	
Final abbreviation	A	B	L	H					

The table for the eliminated letters uses A = 1, . . . , Z = 26; but the table for the derivation of the check letters uses values A = 0, . . . , Z = 25 since it is possible to obtain the value zero if the first number exactly equals a multiple of 26.

Shuffle and Truncation. Shuffle the letters of the word, then drop the letters from the right end until the desired word length is obtained. The shuffle consists of a simple folding of the letters within the word, as in the following example, in which the word ABLATION is reduced to four letters. Different types of folding patterns can be used.



Elimination of Miscellaneous Redundancies. Some benefit may result from operations like the deletion of U after Q, or of one of a pair of the same letters (e.g., tt, mm, nn), or of the second vowel of a double vowel (e.g., ae, io, oe). These possibilities for reduction are not frequent enough to be the basis of a complete technique, but taking advantage of them whenever they do occur can enhance other techniques.

Arithmetic Manipulations. In some types of computer equipment it is possible to do such things as squaring the original word and using some of the digits of the result as the abbreviated word. There would be no resemblance to the original word, but the transformation might yield a relatively unique expression. The abbreviations would probably be numbers, or mixed alphanumeric and special characters. Other types of arithmetic operations could also be used.

Soundex Code. The conventional Soundex system for manual files converts a name to a code word having one alphabetic character and three numeric characters.[12] It is possible to generate larger code

[14] Check digits are generated and used in many ways with data processing systems. However, the application of check digits to word abbreviation to represent the omitted letters is relatively new. (See H. P. Luhn, Chapt. 23 in Casey et al., mentioned earlier in this chapter, fn. 9.)

[12] Remington Rand, *Soundex—Foolproof Filing System for Finding Any Name in the File*, Brochure LBV809 (undated).

Remington Rand *Idem Sonans Says It's Legal (Soundex)*, Brochure LBV528 (undated).

n) Phonetic Coding

Most of the methods of shorthand note taking utilize such phonetic coding (without numbers) to reduce the bulk of material to be written. These shorthand methods obtain further condensation by using a family of special graphic symbols to represent the phonetics. These methods are described later in this chapter.

Evaluation

2. *Comparison of Derived Coding Methods*

In most coding situations, the major objective is to condense the words as much as possible while maintaining a maximum amount of discrimination between the words. Success in meeting this objective can be measured more accurately than for any of the other objectives mentioned previously. Tests have been made on several of these techniques by abbreviating the test material to a given number of characters, then counting the number of unique entries that remain after the abbreviation process.[15] In this way the degree of uniqueness was measured for various amounts of abbreviation. These data provide a measure of the performance of each scheme, and can be plotted to form summary curves of the operating characteristics of each abbreviation scheme.

[15] Bourne, C. P., and D. F. Ford, "A Study of Methods for Systematically Abbreviating English Words and Names," *Journal of the Association for Computing Machinery*, Vol. 8, No. 4, pp. 538-552 (October 1961).

In general, none of the tested techniques produced abbreviations that could be transformed back into the original words, and none of the techniques produced abbreviations that bore a strong resemblance to the original word. With the exception of the statistical data describing the nature of the words to be abbreviated, no table look-up or prior knowledge of the input data was required.

The familiar techniques of simple truncation and vowel elimination produced the poorest results. Schemes that took advantage of the statistical nature of the words (i.e., drop-out by a composite frequency distribution, drop-out by a separate frequency distribution for each letter position, and drop-out by bigram rankings) produced better results. The shuffle scheme produced mediocre results, and the technique of selectively dropping the *n*th letter produced some good results and some bad results. In the two basic schemes (truncation, and selective drop-out of every second letter) that were run with and without the generation of a check letter, the check letter improved the performance.

For subject words, several of the performance curves crossed each other, so that no single scheme was "best," i.e., the "best" scheme varied with the number of letters allowed in the abbreviation. Be-

and an allowable field size of 3 to 6 characters, the best performance was obtained by dropping every other letter and generating a check letter. The test results for proper names did not intersect on the chart and thus there was a single "best" technique for all degrees of abbreviation. Acceptance of these conclusions should be tempered by recognition that not all of the described techniques have been tested.

The subject words of the data base ranged in length from 1 to 23 characters. However, with

The best technique was that of dropping every other letter and generating a check letter, i.e., selective drop-out [illegible] a check letter.

most of the abbreviation techniques, the data could still be represented by 8 to 10 characters while retaining the same degree of discrimination as the original words, i.e., the number of unique file items remained the same after abbreviation even though the field size had been restricted to 8 or 10 characters. The proper names of the data base ranged from 6 to 22 characters after editing. However, these names could also be reduced to 8 to 10 characters while retaining a high percentage of their original uniqueness. These results would probably vary with the total size and configuration of the data base being processed.

ASSIGNED CODES

C. Assigned Coding Methods

Assigned coding methods require the use of detailed code tables for both the coding and decoding operations. Any type of material can be coded by this technique; single words or even large groups of words can be condensed into a concise table entry. For example, in 1874 Bloomer published a code dictionary of frequently used sentences and statements, to reduce the cost of cabling and to provide a certain degree of message security.[16] Examples of his codes are given below, with their assigned meaning:

Filament: Panic in the market. If you want to sell, telegraph immediately.

Phelan: Be careful in dealing with them, as they will take advantage, although their means are ample.

Hoop: Remit to us in gold coin by express, letting portions come daily until all is sent; limiting shipments to the amout of

A similar code dictionary was published in 1913,[17] which included the following:

Loranthus: Sell out at once; do not delay.

Coloring: Do not decide until you receive my letter.

Adjunct: Have had no letters from you for a fortnight or more. Are all well?

Reversible: We all unite in wishing you a very Merry Christmas.

Commercial, government, and amateur radio and teletype operators use a variety of standard code terms to save transmission time and effort. Examples of these code terms are:

GM	Good morning
CUL	See you later
BCNU	I'll be seeing you
73	Best regards
QRM	Are you being interfered with? I am interfered with.
QRU	Have you anything for me? I have nothing for you.
XYL	(Ex young lady) Wife or married woman operator

Some military organizations use standardized abbreviated writing for applications where data must be recorded or transmitted as quickly as possible.[18] An example of this type of notation is given by the following:

POW states he was a Msgr with the 22nd M Tk Bn of the 316th Fus Mecz Div. CO of the Div is Col Hans Heiner, a tough and energetic Off who has written several FMs on Armd tactics and Sup. Div was formed from the remnants of the old 41st Med Tk Brig under a new TOE effective 1 Jan 1954, and was a Repl unit until recently committed to the line. Elms of Svc units from the FIRST Fus Rif Army (Reinf) have been Atchd.

The message is expanded here to illustrate the degree of compression that was obtained.

The prisoner of war states that he was a messenger with the 22nd Medium Tank Battalion of the 316th Fusilier Mechanized Division. The commanding officer of the Division is Col. Hans Heiner, a tough and energetic officer who has written several field manuals on armored tactics and supply. The Division was formed from the remnants of the old 41st Medium Tank Brigade under a new table of organization effective 1 January 1954, and was a replacement unit until recently committed to the line. Elements of service units from the 1st Fusilier Rifle Army (reinforced) have been attached.

This type of abbreviation can easily be generated and interpreted after practice. The message can be filed in abbreviated form and later expanded into the original text.

[16] Bloomer, J. G., *Bloomer's Commercial Cryptograph—A Teletype and Double Index-Holocryptic Cipher* (A. Roman & Co., San Francisco, California, 1874).

[17] ——, *The Adams Cable Codex*, 10th ed. (F. O. Houghton & Co., Boston, 1913).

[18] *Military Symbols*, Dept. of the Army Field Manual FM 21-30 (June 1951).

2. Journal Titles Codes

In the library field, increasing use is being made of a collection of standardized abbreviated notations, called CODENS, to describe the names of magazines, technical journals, conference proceedings, and other periodical publications. The original definition of CODEN was, "a noun meaning the combination of letters, numbers, and symbols assigned to a document or other item as a result of applying a set of coding rules." The first set of codes were suggested in 1953 and consisted of code terms of four alphabetic characters for each periodical name, and usually bearing some mnemonic relation to the publication title.[1] (*Bishop, C., "Coden versus Sigils," AMDO, Vol. 5, p. 28 (1954); Bishop, C., "Use of the Coden System by the Individual Research Scientist," AMDO, Vol. 8, No. ?, pp. 221-226 (1957); Bishop, C., "An Integrated Approach to the Documentation Problem," AMDO, Vol. ?, p. 54 (around 1953)). The intent was to conserve time and space in the handling of serial references. Examples of a few CODENS developed at that time are:

LIBJ	Library Journal
SPLB	Special Libraries
AMDO	American Documentation
LRTS	Library Resources and Technical Services
IEEW	Institute of Electrical and Electronics Engineers. Transactions on Engineering Writing and Speech

The codes are now used by several organizations. Chemical Abstracts Service in particular uses them for a number of their regular publications such as Chemical-Biological Activities. The CODEN have been made an ASTM standard and have been submitted to the American Standards Association for approval as an American standard. They are also under study by several international

organizations. ~~They are used regularly by several organizations and publications including Chemical Abstracts Service.~~

The potential use of the CODEN is felt to be important enough that they are used throughout this book for emphasis and demonstration. A full list of all the CODENS used in this book is given in Appendix ___.

At the time that the CODENS were introduced, Bishop also suggested that a concise code could be developed for the entire periodical citation. For example, an article appearing in the Journal of Biological Chemistry, Volume 195, Number 2, page 663, April 1953, would be coded JBCA-195-663. Chemical-Biological Activities has used this abbreviated citation form for some time, but with more bibliographic elements than Bishop initially suggested. A current CBAC example would be AGVI-A,15,3,65,428-32 which corresponds to Arch. Ges. Virusforsch. Vol. 15, Issue No. 3, 1965, pp. 428-32.

The code collection was initially developed and ~~ascer~~maintained by Bishop, who then transferred the responsibility to the American Society for Testing and Materials (ASTM) in 1961 when the collection had grown to about 4000 codes. In 1963 the code list had grown to 8,500 CODENS and the ASTM published the first directory.[20] (*American Society for Testing and Materials, CODEN for Periodical Titles: An Aid to Storage and Retrieval of Information and to Communication Involving Journal References, T. E. Kuentzel, Ed., ASTM Special Technical Publication No. 329, Philadelphia, Pa. (April 1963).) A 1964 Supplement added an additional 6500 CODEN.[21] (*ASTM, Supplement I, ASTM Special Technical Publication Number 329-S1 (April 1964).) At this time it became apparent that the capacity of the four-character code would be exceeded, and a fifth character was added. The ASTM has continued to

fn 20

fn 21

accept requests for the assignment of new CODENS for periodicals not already included in the directories, and the collection has been growing since then. ~~Approximately 40,000 codes will be~~ a total of 38,993 five-letter codes for periodical titles were included in the 1966 directory.[22] (*add ref.) These codes represent an improvement over some of the abbreviations that have been used somewhat arbitrarily in the literature and cursed thereafter by librarians and scholars[23] (*Richmond, P.A., "Misery is a Short Footnote," LRTS, Vol. 9, No. 2, pp. 221-224, (Spring 1965); and Dunkin, P.S., "Happiness is a Long Footnote," LRTS, Vol. 7, No. 4, pp. 403-405 (Fall 1963).

22. American Society for Testing and Materials. Coden for Periodical Titles, 2nd ed. Philadelphia, Pa., 1966, I. E. Kuentzel, Ed. 2 vol. The fifth character was an A for all entries in this edition. New codes are being added at the rate of about 1000 new codes per month. *(cf. Kuentzel, L.E., "Current Status of the CODEN Project," SPLB, Vol. 57, No. 6, pp. 404-406 (July-Aug. 1966)

3. Book Identification Codes

A considerable reduction in text length, and a convenience in record handling, especially for book ordering and acquisition, could be achieved if each published book could be represented by a unique and nationally recognized identification number instead of using all of the author, title, and other descriptive cataloging information. The Library of Congress catalog card number provides this type of abbreviated representation for some books, but because of the way in which this number is assigned this could not be used for a national book numbering system. Previous studies have indicated the desirability of developing such a number, and have made specific proposals for a uniform national numbering system.* (*Harrer, G. A. and H. Ladenson, "A Proposal for a National Code Number System for Current Publications," LRTS, Vol. 6, No. 1, pp. 4-12 (Winter 1962); McCutchen, C. W., "Random Code Numbers for Universal Identification of Documents," AMDO, Vol. 16, No. 2, pp. 91-96 (April 1965); and Melcher, D., "Books Without Number," Publisher's Weekly, Vol. ?, No. ?, pp. 24-27 (15 August 1966)). British book publishers have been implementing such a numbering system since the beginning of 1967,* (*Melcher, D., "British Books to be Numbered Beginning This Month," Publisher's Weekly, Vol. ?, No. ?, pp. 28-30 (2 January 1967)). The numbers will be assigned for publication in the major bibliographies such as British Books in Print and the British National Bibliography and for use by the booksellers and libraries. In Britain, the number previously used by the British National Bibliography (a number analagous to the Library of Congress Card Number) will be dropped and replaced by the new number. This means that a single uniform identification number will be used for books that are published in Britain. The British number is built around

a 9-digit number in three distinguishable parts: a publisher identification number, a serial number for the books of the corresponding publisher, and a check character.* (*"Standard Book Numbering, a report of the Standard Book Numbering Agency, London. 1967. SBN 949999008). The first part (called the publisher prefix) may be two to seven digits in length and always identifies the publisher. The larger the publisher ~~item~~ than the smaller number of digits used to identify him. In this way there are more digits available in the second part of the number. The second part (the title number) of an SBN identifies the title, or a particular volume or edition if there is more than one volume or edtion of the title concerned. Title numbers can be one to six digits in length, and their length is governed by the number of digits in the publisher prefix that precedes them. The publisher prefix and the title number always total eight digits. A ninth digit is used for a check digit, and is computed in a manner described in the section of this book on check digits.

Examples of these Standard Book Numbers are:

SBN 8436 1072 7

SBN 949999 00 8

U. S. publishers are presently implementing a similar system.

Foster, F.G., "Standard Numbering in the Book Trade," the Publishers Association, London, 1967. 9pp.

4. Subject Classification Codes

Most classification schemes that are developed to organize a body of information into some logical groupings are usually assembled and printed with the full English spelling given for the various groupings and sub-groupings. However, while using the scheme it is often very convenient to assign some shorthand representation or code to each of the ~~English~~ text terms that describes and defines the framework of the classification system. Common examples are the familiar library ~~description~~ classification systems of Dewey Decimal and Library of Congress codes. The use of the shorthand notation makes it easier to record the classification headings in a more compact form. The abbreviated notation may also permit some additional conveniences in the machine handling of the subject headings. A distinction should be made between the classification scheme itself, and its system of notation or coding. For any classification scheme, a variety of notations could be used, and each would have little effect on how well the scheme organized a collection of information. The notation used for a particular scheme is a matter that should be considered independently of the basic question of how a body of information should be grouped and organized. Unfortunately, classification systems are often known and criticized primarily on the basis of their notation, even though it would be quite possible to completely change the system of notation without changing the basic structure of the classification scheme.

If the indexing system is a coordinate indexing system in which no index terms are subordinate to any other terms, then there is no need to build a generic structure into the code. The ASTIA Thesaurus Code Manual,

provides one example of such a code, and consists of a table of code numbers for each descriptor used in their indexing system.[24] ([26] ASTIA Thesaurus Code Manual, an un-numbered report of the Armed Services Technical Information Agency (June 1961). The use of fixed-length numeric codes numbers instead of the full alphabetic spelling of the descriptors simplifies the subsequent computer handling of the index records. A seven-digit number is assigned to each descriptor so that the numeric ~~sequence~~ and alphabetic sequences coincide. Examples of their codes are:

fn 26

astronomy	0534000
beeswax	0711000
cats	1098000

This type of fixed length numeric code provides room for expansion by leaving vacant numbers between the ones that are used. However, there is a definite limit to the number of items that can be included in this restricted code field.

Numeric codes can also be constructed so that they contain a generic or heirarchical structure that permits unlimited expansion by subdivision. This is accomplished at the expense of a greater number of digits. One of the most common library examples is the Dewey Decimal System in which further subdivisions of a main topic can be distinguished by further partitioning and detail branching of the number sequence. Consider the following selected example from this system:

The users of most general or public libraries do not need the great detail of subject subdivision that can be provided by such a code, consequently these libraries often restrict or truncate the code to three or four significant digits. However, other libraries with large collections in very specific subject fields may have users that require a more detailed subject division, and thus may use all of the code resolution or detail that is available ~~furnished~~. They may even develop some of their own more detailed subject subdivisions and codes to use within the more general framework of the available classification schedules. These special collections will generally have longer code terms than the general libraries. For example, 621.312 136 is the Dewey Decimal notation for wind-driven electric plants. The Universal Decimal Classification (UDC) is another numeric notation system with a structure similar to the Dewey system. For example, 551.507.362.2 represents artificial satellites in the UDC system. The similarity stems from the fact that they both developed as offshoots from the same parent system. An example of selected portions of the UDC schedule is given ~~below~~ in Table 3-7 in order to illustrate the notation and generic structure.

Mixed alphanumeric codes have also been used. The Library of Congress classification system is the best known example of this type of notation. The alphabetic prefix represents a general subject area, and the accompanying numeric code represents a more detailed division of that subject area. A complete outline of the LC subject classes was given ~~in Table~~ earlier in Chapter 2. An example of the specific type of notations used within a subclass is given below with a few codes selected from the general schedule.

FIGURE II

FORM AND CONTENT FACETS IN UDC
WITH STANDARD NOTATION

Facets		Standard Notation
Form facets:		
1.	language	=...
2.	form of work	(0...)
Content facets:		
3.	place	(n...) n=1 to 9
4.	race	(=...)
5.	time	"..."
6.	general subject	absence of notational signal

FIGURE III

EXAMPLE OF UDC INDEXING OF A DOCUMENT

Document Title: Distribution and Seasonal Movements of Saginaw Bay Fishes.

UDC in statement form:

597:591.9:591.52(285:71:73)"1964/1966"(047+084.3)

UDC in index entry form with English equivalents:

(047)	Technical Reports
(084.3)	Maps
(285)	Lakes
(71)	Canada
(73)	United States
"1964/1966"	Events of 1964 - 1966
591.52	Animal Habitats and Migrations
591.9	Geographical Distribution of Animals
597	Ichthyology. Fish.

ABSTRACT

Freeman, Robert R. and Pauline Atherton, File Organization and Search Strategy Using the Universal Decimal Classification in Mechanized Reference Retrieval Systems, Report AIP/UDC-5 under National Science Foundation Grant GN-433, New York, American Institute of Physics, September 15, 1967.

Starting from a model of contemporary mechanized retrieval systems and the characteristics of indexing languages used therein, the authors develop a rational basis for use of the Universal Decimal Classification (UDC) in this context. Practical design considerations for the use of UDC in a mechanized retrieval system are discussed. Examples are reported of the use of UDC as the indexing language with the Combined File Search System, an existing retrieval system for the IBM 1401, used by several large information centers in the United States. Finally, the authors discuss how UDC might be used as a query language in a typical retrieval system of the near future in which the user interacts directly with the computer-stored document reference file.

The authors conclude that it is technically feasible to use UDC in mechanized retrieval systems and that, under certain conditions, it may be desirable. Some of these conditions are the existence of large files already indexed by UDC, staff already trained for its use, and extensive international use or exchange of materials of the system.

UDC 025.3+025.4UDC:651.53+651.838.8:681.322.04/.06

Explanation of UDC numbers:

025.3	–	Cataloging and indexing - Information retrieval systems
025.4UDC	–	Decimal classifications - U.D.C.
651.53	–	Methods of file organization
651.838.8	–	Electronic methods of selection from files
681.322.04/.06	–	Digital computers - data representation (.04), data conversion (.05), and programming aspects (.06)

625	Railway and Highway Engineering
625.01	Parts of Vehicles except Propelling Engines
625.02	Principles of Road Surface Construction
625.021	The Principle of Macadam Pavement
625.022	The Principle of Cement Concrete Pavement
625.023	The Principle of Mortar
625.032	Chemical and Physical (mainly mechanical) Forces and Equilibrium of Forces between Track and Vehicle
625.032.3	Forces and Equilibrium of Forces in Contact Between Track and Single Wheel
625.032.32	Surface Properties of Track and Wheel. Roughness, etc.
625.033	Stresses set up in Vehicles and Track
625.033.3	Stress in Track (either level or inclined)
625.033.32	Static Stress on Straight Track
625.033.37	Track Deformation. Displacement of Track
625.033.372	Through Fissures and Flaws in Track or Road Surface
625.033.373	Through Wear of Rails
625.033.373.2	Normal Wear
625.033.373.4	Wear Arising from Braking

Table 3-7

Example of ~~Outline of Library of Congress~~ Selected Portion of UDC Subject Classification

H Social Sciences

HD Economic History

7513 Public Buildings

.A3 General Works

7515 Department Stores

S Agriculture--Plant and Animal Industry

T Technology

TL Aeronautics

722 Use of Airplane in Various Industries

.1 Agriculture

.2 Fisheries and Whaling

723 Government uses of airplanes

724 General

U Military Science

There are also many examples of subject codes constructed entirely from alphabetic characters. Combination or faceted codes that are composites of several different classification systems often include a notation mechanism (e.g., a colon or other special character) to separate the codes. An example of such a code notation is given below.

L2153:4725:63129:B28 = soft palate:cancer:radium treatment:statistical study

L = medicine

L2 = digestive system

L21 = mouth

L215 = palate

L2153 = soft palate

L2153:4 disease

and so on. This and other interesting examples of subject code notations

are given in two IBM publications.[27] (*IBM Corp., "Reference Manual: Index Organization for Information Retrieval," 1961. Report C20-8062; and "Modern Coding Methods," undated. Report X21-3793-6). Other good examples of codes and notation are given in books by Kent.[28] (*Kent, A., "Textbook on Mechanized Information Retrieval," Interscience Publishers, N.Y. 1962) and by Becker and Hayes.[29] (*Becker, J. and R. Hayes, "Information Storage and Retrieval: Tools, Elements, Theories," Wiley, N.Y. 1963).

The UDC uses a variety of notation tricks.[30] (*International Federation for Documentation, "Universal Decimal Classification," 2nd ed. Rev. 1957. London. British Standards Institution. 1957.) The subjects are described by numbers (e.g., 2 = religion, theology; 5 = pure sciences) as noted earlier in chapter 2. ~~Table ___. It uses~~ a plus (+) sign is used to show subjects covered jointly in a book (e.g., 2+5 = religion and pure science). ~~It uses~~ a slash (/) sign is used to denote a run of consecutive subject numbers (e.g., 71/75 = 71+72+73+74+75). ~~It uses~~ a colon (:) sign is used to show a relationship. ~~It uses~~ a variety of other tricks are used to help make the final notation more precise and more flexible. Suggestions have been made to increase the capacity of such a notation system even further by using a number base other than decimal.[31] (*Richmond, P. A., "A Simple Mnemonic Large-Base Number System for Notational Purposes," JDOC, Vol. 14, No. 4, pp. 208-211 (December 1958).)

5. Chemical Coding

Since 1861, chemists throughout the world have been using and developing methods for describing molecular structures in a linear or line-formula notation (printable pictures). Many notation systems have now been developed which provide a concise representation of a chemical compound, and permit some searching and manipulation of such lists of information.*

NAS-NRC Committee, "Survey of Chemical Notation Systems," Report 1150, National Academy of Sciences - National Research Council, 1964.

Frear, D. E. H., Survey of European Non-Conventional Chemical Notation Systems. National Academy of Sciences - National Research Council. Washington, D.C. Publication 1278, 1965, 78 pp.

Bonnett, H. T., "Chemical Notations -- A Brief Review," JCHD, Vol. 3, No. 4, pp. 235-242 (October 1963).

Hayward, H. W., "A Second Look at Chemical Notation Systems in View of Projected Machine Interconversion of Cipher Forms," PADO, Pt. 1, pp. 59-60. (American Documentation Institute, Washington, D.C., 1963).

Huber, M. L., "Chemical Structure Codes in Perspective," JCHD, Vol. 5, No. 1, pp. 4-8. (Feb. 1965).

This becomes increasingly important now as an estimated 28,000 new chemical compounds are made each year, to add to the estimated total of two million unique organic compounds already developed and reported.*

Wiswesser, W. J., "A Decoding Study of the Line-Formula ~~(or?)~~ Chemical Notation," December 1965. Reading Chemists' Club, Reading, Pa.

The notation schemes typically use a few new symbols to concisely and unambiguously denote the commonly encountered structure segments.

For example,

diethylene glycol, $OH-C_2H_4-O-C_2H_4-OH$, is denoted Q202Q

where the Q is used to denote the OH-group, and the numerals denote the number of carbon atoms in alkyl (ene) chains. One example of the way in which such a notation permits further processing and accessing from multiple points is the WKIC index to chemical notations, as illustrated by the examples in Figs. 3-4.*

Sorter, P. F., et.al., "Rapid Structure Searches via Permuted Chemical Line Notations," JCHD, Vol. 4, No. 1, pp. 56-60, (January 1964).

Granito, C. E., et.al., "Rapid Structure Searches via Permuted Chemical Line Notations, II. A Key Punch Procedure for the Generation of an Index for a Small File," JCHD, Vol. 5, No. 1, pp. 52-55 (February 1965).

Horner, J. K., "Low Cost Storage and Retrieval of Organic Structures by Permuted Line Notations: Small Collections." JCHD, Vol. 7, No. 2, pp. 85-88 (May 1967).

Granito, C. E., et.al., "Rapid Structure Searches via Permuted Chemical Line Notations, A Computer-Produced Index," JCHD, Vol. 5, No. 4, pp. 229-233 (November 1965).

Gelberg, A. "Rapid Structure Searches via Permuted Chemical Line Notations: IV. A Reactant Index," JCHD, Vol. 6, No. 1, pp. 60-61 (Feb. 1966).

A similar permuted index approach is taken by the Rotated Formula (Rotaform) Index to Index Chemicus as illustrated in Figs. 3-5. This permuted index forms all possible written sequences of elements for a given compound, and lists each compound under each of the several formulas that might be used to describe it.*

Garfield, E., "Generic Searching by Use of Rotated Formula Indexes," JCHD, Vol. 3, No. 2, pp. 97-103, (April 1963).

A straight alphabetic molecular formula index is sorted and printed

for Chemical-Biological Activities (see Fig. 3-6).

A number of named notation systems are in use today, including the Wiswesser,*

Wiswesser, W. J., Sorter, P. F., and Granito, C. E., previous citations.

Wiswesser, W. J., "A Line-Formula Chemical Notation." Thomas Y. Crowell Co., New York, 1954.

Gelberg, A., "A Program Retrieval of Organic Structure Information via Punched Cards," JCHD, Vol. 2, No. 1, pp. 7-11 (January 1962).

Bonnett, H. T., and D. W. Calhoun, "Applications of a Line Formula Notation in an Index of Chemical Structions," JCHD, Vol. 2, No. 1, pp. 2-6 (January 1962).

Smith, G. E., "Machine Searching for Chemical Structures," SCIE, Vol. 131, pp. 142-146 (January 15, 1960).

Bowman, C. M., et. al., "A Chemically Oriented Information Storage and Retrieval System. I. Storage and Verification of Structural Information," JCHD, Vol. 7, No. 1, pp. 43-47 (February 1967).

Granito, C. E., et. al., "Use of the Wiswesser Line Notation for Determining Duplicate Chemical Structures," JCHD, Vol. 6, No. 4, pp. 252-253 (November 1966).

Wiselogle,*

Buhle, E. L., et. al., "A New System for the Classification of Compounds," JCED, Vol. 23, pp. 375-391 (1946)

Arendell, F. H., "A Three-Symbol Code for Searching Chemical Structures," JCHD, Vol. 1, No. 3, pp. 47-57 (1961).

CBCC,*

Seitner, P. B., "Biology Code of the Chemical-Biological Coordination Center. Key to the Biology code of the Chemical-Biological Coordination Center," Publications 790 and 790K, National Academy of Sciences -- National Research Council, Washington, D.C. (1960).

Welt, I. D., "Aspects of the CBCC Biology Code of Interest to Chemists," JCHD, Vol. 1, No. 1, pp. 19-21 (1961).

Craig, P. N., "Correlation of Chemical Structures and Biological Data," JCHD, Vol. 2, No. 2, pp. 106-109 (April 1962).

Dyson,*

Dyson, G. M., "A New Notation and Enumeration System for Organic Compounds," 2nd ed., Longmans, Green and Co., New York (1949).

Dyson, G. M., and E. F. Riley, "Use of Machine Methods at Chemical Abstracts Service," JCHD, Vol. 2, No. 1, pp. 19-22 (January 1962).

and Hayward.*

Hayward, H. W., et. al., "Some Experience with the Hayward Linear Notation System," JCHD. Vol. 5, No. 3, pp. 183-189 (August 1965).

Hayward, H. W., "A New Sequential Enumeration and Line Formula Notation System for Organic Compounds," Research and Development Report No. 21 of the U. S. Patent Office. November 1961.

Marron, B. A., et. al., "Chemical Substructure Searching with Linear Notations," JCHD, Vol. 6, No. 2, pp. 92-95 (May 1966).

Many other codes have been suggested and used, and references to many of these are listed at the end of this chapter.

In addition to the conventional molecular formula and some form of linear notation, a chemical compound can also be specified by a systematic chemical name or nomenclature, a structural diagram, a unique Chemical Abstracts Service registry number, and a connection table. The chemical structure description used in the Chemical Abstracts Service Registry is a uniquely ordered list of the node symbols of the structure (or graph) in which the value (atomic symbol) of each node and its attachment (bonding) to the other nodes of the total structure are described. Such a list and description is called a "connection table."*

Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures -- A Technique Developed at Chemical Abstracts Services," JCHD, Vol. 5, No. 2, pp. 107-113 (May 1965).

Penny, R. H., "A Connectivity Code for Use in Describing Chemical Structures," JCHD, Vol. 5, No. 2, pp. 113-117 (May 1965).

Gluck, D. J., "A Chemical Structure Storage and Search System Developed at Du Pont," JCHD, Vol. 5, No. 1, pp. 43-51 (February 1965).

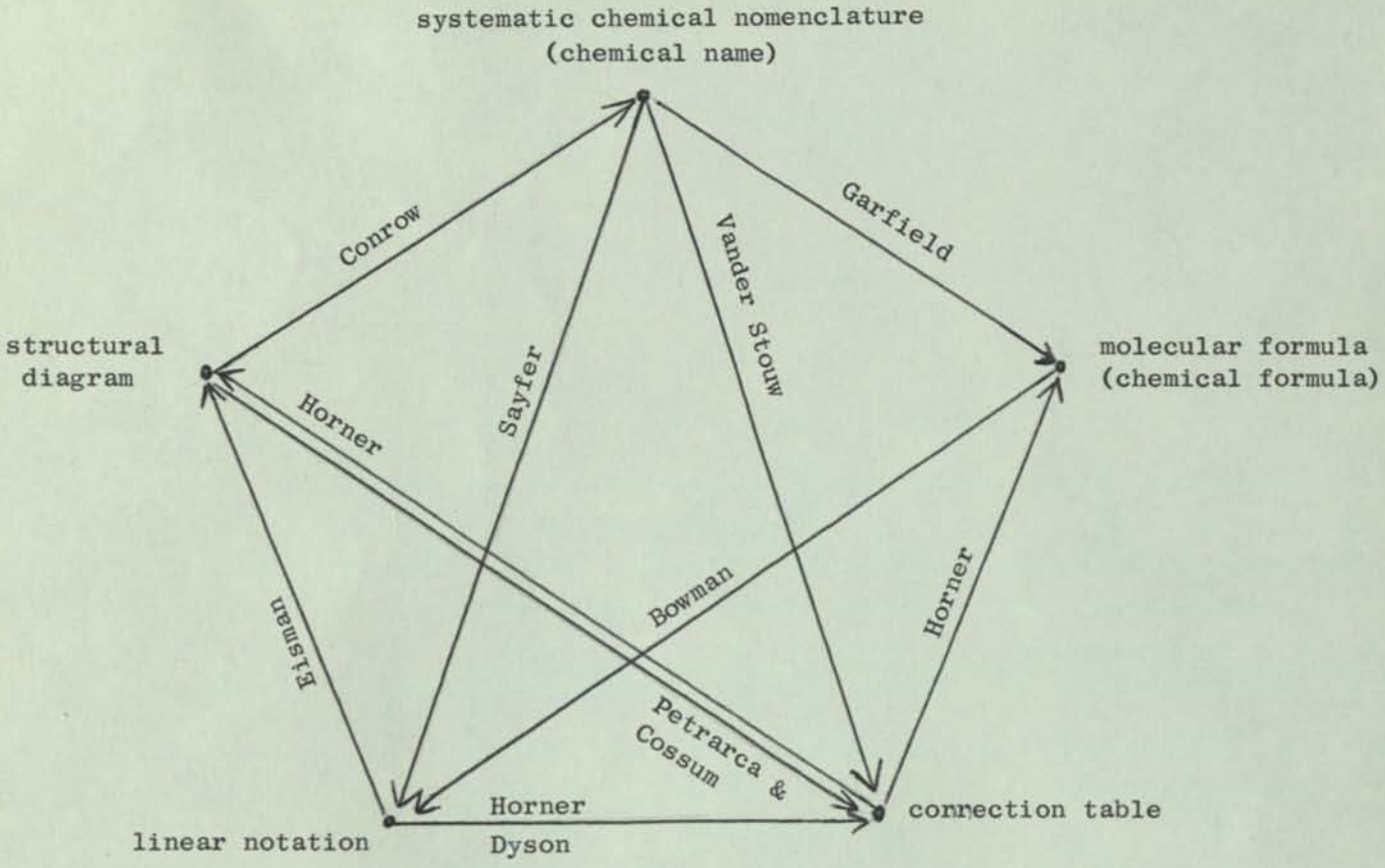
Spialter, L., "The Atom Connectivity Matrix (ACM) and Its Characteristic Polynomial (ACMCP)," JCHD, Vol. 4, No. 4, pp. 261-269 (October 1964).

Rules and procedures are now available to convert automatically from some of these representations into others (e.g., from linear notation to a structural diagram). An indication of the degree to which these automatic procedures have been developed is given by the illustration in Fig. .

In addition to coding of specific chemical compounds, efforts have also been made to develop codes for smaller groups of compounds or mixtures of special interest. For example, the Food and Drug Administration has recently established a National Drug Identification Number to uniquely describe each different drug entity, with a separate identification number for each combination of manufacturer/product name/dosage form/dosage strength. Examples of such codes are given below.

Over numbers have been assigned to date by FDA, and this identification system continues to grow in importance to pharmacuetical manufacturers, pharmacists, hospital administrators, medical insurance carriers, and the medical profession. The American Society of Hospital Pharmacists is working on a computer-based numerical code for major drug products marketed by American manufacturers. Directories of drug codes have also been prepared by several state and federal agencies.

Chemical Abstracts Service has developed a Chemical Compound Registry Number of five to nine digits, and a Registry system, to permit the assign-



Note: 1. The direction of the transformation is shown by the arrowhead.

2. The names associated with each link refer to the references given below.

Bowman, C.M. et. al., "A Chemically Oriented Information Storage and Retrieval System I. Storage and Verification of Structural Information," JCHD, Vol. 7, No. 1, pp. 43-47 (February 1967).

Conrow, K., "Computer Generation of Baeyer System Names of Saturated Bridged Bicyclic, Tricyclic, and Tetracyclic Hydrocarbons," JCHD, Vol. 6, No. 4, pp. 206-213 (November 1966).

Cossum, W.E., and M.E. Hardenbrook, "Computer Generation of Atom-bond Connection Tables from Hand-drawn Chemical Structures," PADO. 1964. pp. 269-275.

Dyson, G.M., " ? " ISSR, Vol. ?, No. ?, pp. ? (? 1963).

Fig.

Representative Procedures to Transform automatically from one Chemical Representation to Another

Eisman, S.H., "A Polish-Type Notation for Chemical Structures," JCHD, Vol. 4, No. 3, pp. 186-190 (July 1964).

Garfield, E. "An Algorithm for Translating Chemical Names to Molecular Formulas" JCHD, Vol. 2, No. 3, pp. 177-179 (July 1962).

Horner, J.K. " ? "

Petrarca, A.E., "A Method for Generating Unique Computer Structural Representations of Stereoisomers," JCHD, Vol. 7, No. 3, pp. 154-165 (August 1967).

Vander Stouw, G.G., et. al., "Procedures for Converting Systematic Names of Organic Compounds Into Atom-Bond Connection Tables," JCHD, Vol. 7, No. 3, pp. 165-169 (August 1967).

ment of a unique identification number to each chemical compound that is new to the Registry file. The Chemical Compound Registry, under development at Chemical Abstracts Service since 1961, is a structure-based computer store of compound-oriented information. It is hoped that eventually it will include data on all publicly reported chemical compounds. The system may be searched through the chemical name, the Registry number, or the chemical structure, and all data on that compound will be retrieved. Since January 1965, all compounds indexed for Chemical Abstracts have been registered. By late 1967, this included 600,000 unique structures with 800,000 associated names and 1.2 million bibliographic references. Compounds new to the file are being added at a rate of about 4000 per week from the current literature. Approximately 3 to 4 million compounds remain to be included in the file. These Registry numbers are now starting to be cited with the compound descriptions in some of the journal articles, and will continue to receive increased use in the future.*

Leiter, D. P., et. al., "Installation and Operation of a Registry for Chemical Compounds," JCHD. Vol. 5, No. 4, pp. 238-242 (November 1965).

VOL PAGE	* REFERENCE POINT	NAME
III286	2OVYCN&R BQ EQ DYCN&VO2	2,5-DIOL-1,4-BENZENE BIS-CARBETHOXYA CETONITRILE
IV 953	NCYCN&YCN&R DN1&1	1,1,2-TRICYANO-2-(4-DIMETHYLAMINOPHE NYL)ETHYLENE
IV 776	NC2XCN&R&VO2	ETHYL 2,4-DICYANO-2-PHENYLBUTYRATE
IV 877	NCYCN&UYCN&CN	TETRACYANOETHYLENE
IV 463	2OVYCN&UY1&R	ETHYL (1-PHENYLETHYLIDENE)CYANOACETA TE
IV 93	2OVYCN&UY2&1	ETHYL 2-CYANO-3-METHYL-2-PENTENOATE
IV 515	NCYCN&U1MVM7	3-HEPTYLUREIDOMETHYLENEMALONONITRILE
IV 566	2OVYCN&U1O2	ETHYL ETHOXYMETHYLENECYANOACETATE
I 181	QVYCN&U1R	2-CYANO-3-PHENYLACRYLIC ACID
III397	2UY2&X1&CN&VO2	ETHYL 2-CYANO-3-ETHYL-2-METHYL-3-PEN TENOATE
III399	2Y2&UYCN&VO2	ETHYL 2-CYANO-3-ETHYL-2-PENTENOATE
IV 97	3X2&1&YCN&VO2	2-CYANO-3-METHYL-3-ETHYLHEXANOATE
III286	2OVYCN&R BQ EQ DYCN&VO2	2,5-DIOL-1,4-BENZENE BIS-CARBETHOXYA CETONITRILE
IV 174	GR DYCN&V1R	2-(4-CHLOROPHENYL)-4-PHENYLACETOACET ONITRILE
IV 175	ER DYCN&V1R	2-PHENYL-4-(4-BROMOPHENYL)ACETOACETO NITRILE
IV 175	1R DYCN&V1R	2-PHENYL-4-(4-TOLYL)ACETOACETONITRIL E
IV 175	1OR DYCN&V1R	2-PHENYL-4-(4-ANISYL)ACETOACETONITRI LE
IV 175	1SR DYCN&V1R	2-PHENYL-4-(4-METHYLTHIOPHENYL)ACETO ACETONITRILE
IV 97	2OVYCN&X2&1&R	2-CYANO-3-METHYL-3-PHENYLPENTANOATE
IV 93	2OVYCN&X2&1&1R	ETHYL 2-CYANO-3-METHYL-3-BENZYLPENTA NOATE
IV 953	NCYCN&YCN&R DN1&1	1,1,2-TRICYANO-2-(4-DIMETHYLAMINOPHE NYL)ETHYLENE
III59	ZVYCN&Y1&YVZCN	2,2-DICYANO-3-METHYLGLUTARAMIDE
I 20	U2NCN&2U1	DIALLYLCYANAMIDE
I 44	SCNR	PHENYL ISOTHIOCYANATE
IV 70	SCNR	PHENYL ISOTHIOCYANATE
IV 70	SCNR BG	2-CHLOROPHENYL ISOTHIOCYANATE
IV 70	SCNR DE	4-BROMOPHENYL ISOTHIOCYANATE
I 16	SCNR DG	4-CHLOROPHENYL ISOTHIOCYANATE
IV 70	SCNR DR	4-PHENYLYL ISOTHIOCYANATE
III73	SCNVR	PHENYLTHIOCYANATE
III59	SCN1	METHYL ISOTHIOCYANATE
III84	OCN11	UNDECYL ISOCYANATE
III60	SCN2	ETHYL ISOTHIOCYANATE
IV 43	CN2	ETHYL ISOCYANIDE
IV 81	OCN4NCO	BUTANE 1,4-BIS-ISOCYANATE
IV 52	OCN6NCO	HEXAMETHYLENE DIISOCYANATE
IV 81	4NCO	BUTANE 1,4-BIS-ISOCYANATE
IV 52	6NCO	HEXAMETHYLENE DIISOCYANATE
II 45	DNCO	4-NITROPHENYL ISOCYANATE

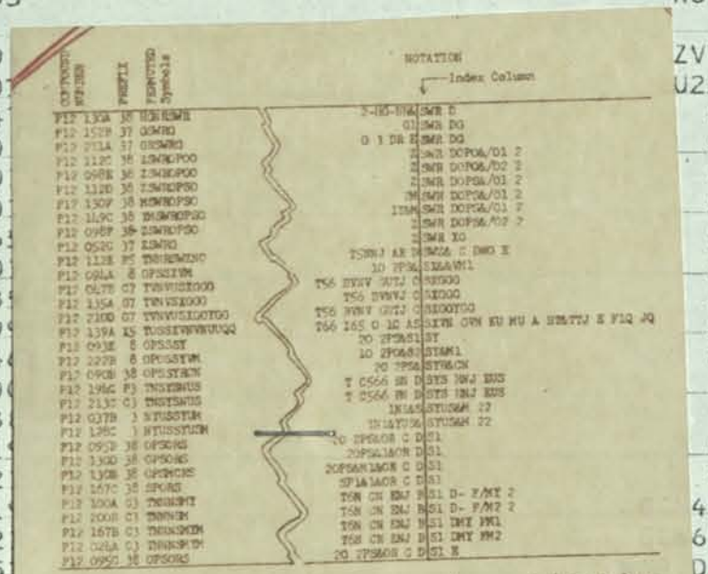

NOTATION
Index Column

Figure 4. Example of printout page of permuted notations for some sulfur compounds appearing in "Pesticide Index" (2nd Ed.). The designations in the Compound Number column refer to the page number appearing in "Pesticide Index."

"Computer Structure Searching" JCHD Vol 5 No. 4 Nov 1965

Sample Portion of a Permuted Index of Wiswesser Notations

provides a reference code for each urban place with a population of over 2500 as determined by the 1960 census.

6. Geographic Place Name Codes

The user with a problem of recording city or other geographical names might consider the use of a simple numeric code for states, counties, and cities. Codes of this type have been used in many data processing systems.[41]*

(*IBM Corp., Numerical Code for States, Counties, and Cities of the United States, ~~Brochure X21-4653~~ Report C20-8073-0 of the IBM Corp., ~~New York (1958))~~ Data Processing Div, White Plains, N.Y. 1961; and "Numerical Code for States, Counties and Cities," Newsletter No. N20-1022-0. 15 May 1967 from the same organization provides some corrections to the original report.) In such a code, ~~the~~ major place names ~~is~~ are often represented by a unique number, and sometimes include~~s~~ a generic structure built into the code pattern. Examples of such codes are:

City	State Code	County Code	Population Group	City code
Cleveland	34	035	12	0510
Columbus	34	049	21	0570
Los Angeles	04	037	11	1900
Minneapolis	22	053	21	0650
Palo Alto	04	085	44	2520
Washington, D.C.	08	001	12	0010

Place names have also been represented by alphabetic codes. An example of such a code is the one used in the present Library of Congress MARC computer catalog records. In this scheme, the place of publication is represented by a four letter mnemonic code. The first two letters of the code represent the country, if foreign; or the state, if within the United States. The last two letters represent the city. This allows a maximum of 676

cities within each county or state to be represented. The code for the city is not unique unless combined with the code for county or state. Examples are:

CAPA	Palo Alto, California
CAMP	Menlo Park, California
VAFX	Fairfax, Virginia
ENLO	London, England
JATO	Tokyo, Japan

Almost all attempts to maintain a code term's resemblance to the original spelling or pronounciation fail when the list grows to more than a few hundred terms.

Another example of area coding is that given by the U.S. Post Office ZIP code for mail addressing and distribution which shows the general region of the country, as well as a specific post office within that region.* (*U.S. Post Office Dept. National ZIP Code Directory, Wash., D.C. January 1966). The lower order digits of the code can be used to establish more detailed ZIP code breakdowns within the jurisdiction of a major post office area. For example, Chicago has a general ZIP code of 60600, but also uses

a range of ZIP codes from 60601 - 60690 to identify more specific regions within Chicago. In some instances a single organization such as a large government office may have its own unique ZIP code number.

20540 Library of Congress

20541 Library of Congress--Card Division

20401 Government Printing Office

94025 Menlo Park, California

Airports throughout this country and the rest of the world are designated for ticketing and scheduling purposes by a universally accepted coding system. The frequent traveler will recognize such familiar codes as:

DIA Dulles International Airport, Washington, D.C.

JFK Kennedy " " New York City

SFO San Francisco " " San Francisco

LAX Los Angeles " " Los Angeles

7. Organizations Codes

Organizations, like places, can be represented by numeric, alphabetic, or mixed alphanumeric codes. Examples of alphabetic codes can be seen in several of the published union lists of library holdings. Some specific examples from the Union Library Catalogue of the Philadelphia Metropolitan Area are:

ColCheHill	Chestnut Hill College
HamWatch	Hamilton Watch Company Library
UofPHyg	University of Pennsylvania Microbiology Library.

As shown above, the codes need not be of uniform length or case. One other well known application of organization codes is that made by the stock exchanges. As in the prior union list examples, the codes are variable length and mixed case:

Amer Motor	American Motors
BorgWr	Borg Warner
ComwEdis	Commonwealth Edison.

The most obvious example of variable length alphabetic coding is our system of abbreviations for state names. See 43

These types of codes have been helpful for simplifying the entry of data or queries into a computer system.[43] (43. Jackson, M., "Mnemonics," DTMN, Vol 13, No. 4, pp. 26-28 April 1967)

Publisher names are variable length coded for the Library of Congress MARC computer catalog records. A two-, three-, or four-letter mnemonic code is used. In some instances of government agencies the code will also designate the country or state, as well as a specific publisher. In other instances a publisher will be assigned a discrete code without a coded indication of his location. Examples of the code are:

GPO	Government Printing Office
USFZ	all agencies of the U.S. federal government
CNKZ	a Connecticut state government agency at the county level
WIL	John Wiley & Sons, Inc.
STUN	Stanford University

There are also numerous examples of fixed length alphabetic codes, such as the standard ones used for reservations and ticketing for all scheduled airlines:

TWA	Trans World Airlines
UA	United Airlines
AA	American Airlines

Our modern age has also spawned a large number of abbreviations and acronyms that have received such common use that it is often difficult to

recall exactly the full and original name of the organization. Common examples of organization acronyms in the library field are:

COSATI	Committee on Scientific and Technical Information
CFSTI	Clearinghouse for Federal Scientific and Technical Information
DDC	Defense Documentation Center

Military organizations have also adopted abbreviated organization names, usually constructed in a way that retains a close similarity to the original name. For example:

COMSUBDEVGRU TWO	Combat Submarine Development Group No. 2
CINCPAC	Commander-in-chief, Pacific
COMUSMACV	Commander, U.S. Military Assistance Command, Vietnam

One example of a numeric organization code is the DUNS number (Data Universal Numbering System) developed by Dun & Bradstreet, Inc. to uniquely identify each of over 500,000 companies.[4] (*"Data Universal Numbering System," Dun & Bradstreet, Inc., N.Y., brochure 18A-13(67 03 10). March 1967). Dun & Bradstreet assigns the numbers, advises each company of its number, maintains the master file, and provides subscribers with code books in printed or machine-readable form. DUNS numbers are assigned to "establishments" which are considered to be a single physical location, either

an operating location or a non-operating headquarters address. A company may consist of more than one establishment, each of which will have its own DUNS number. The number is a randomly assigned eight digit number. The last digit of the number is a check digit computed by the "Modulus 10, double-one-double-one" method described later in this chapter (~~see Sec. 2~~). The number is a completely random number and represents one establishment at one location. It has no built-in significance, and is meant to be used only as an identification code.

8. Account Identification Codes

Business, government, and other large organizations ~~have a~~ need ~~for~~ to uniquely identify~~ing~~ individual customer accounts. There are too many people with similar names, too many name changes, and too many address changes to make this a simple task. It was this type of identification problem that led to the development of the Social Security number--a number that would unambiguously denote the account or record of a particular individual. The use and application of this number has expanded since that time, but it was initially intended to be an account identification number. Many organizations with large numbers of accounts have also developed unique code numbers to use in place of names and addresses. One example is the account number (instead of a name and address) stamped on retail credit cards. Another less familiar example is the account code used by many publishers to keep their subscription records in order. Picture the problems of maintaining an up-to-date mailing list of 40 million names!

The address label in Fig. 3-7 represents an actual sample used by ~~a~~ two large publishers for mailing their magazines. Notice that in addition to the name and address there is also a line of coding (sometimes referred to as a keyline or matchcode). A keyline of this type is used on mailing

Fig 3-7

005 4200 BOURSN 1619 C28 Z

MR. CHARLES P. BOURNE

1619 SANTA CRUZ AVE.

MENLO PARK, CALIF. 94025

In the first line of the mailing label:

0 = area code for U.S. (1 = Canada, etc.)

05 = state code

4200 = city code for Menlo Park

BOUR = first four letters of last name

SN = first and third letters of street name (Santa Cruz Ave.)

1619 = house number

C = first initial

Z = the unique character (assigned in descending sequence)

8 = check digit

NOV66 GAN LK850J98L99 06 11
MR JOHN T GRANT
850 LAKE SHORE DR
LAKE SHORE CLUB
CHICAGO IL 60611

In the first line of the mailing label:
NOV 66 = subscription expiration date
GAN = surname (1st, 3rd, and 4th letters)
LK = street name (1st & 3rd letter)
850 = street number (last 3 numbers)
J = first initial
9 = unique character
8 = check digit
L = specific magazine
99 = state code
0611 = zone or city

Fig. 3-7

REPRESENTATIVE KEYLINE ELEMENTS FOR MAILING LABEL

labels for magazine subscription mailing by most publishers. The keyline serves as an account number for uniquely identifying records in the file, and is changed with name and address changes. Since it is sometimes necessary to locate records in the file when the keyline is unknown, it has been found convenient to construct the keyline from several elements of the account identification such as the individual's name and address. ~~All~~ 5 keylines are normally constructed and assigned by ~~the~~ a computer, extracting the

necessary characters and performing lookups in stored reference tables such as postal directories. The principal problem in the generation of keylines by the computer is determining where the street name begins (e.g., N. Main St.; North Main St.; Third St., S.W.) or where the last name beings (e.g., Prof. E. Van der Pelt; Mr. Joe Van Johnson, Ph.D.; Karl Gusti, Prof. Ing.) A ~~The~~ unique character may be ~~is~~ assigned to the keyline in case of ties (i.e., different accounts that yield the same keyline).

The use of this type of a keyline also permits the computer records to be sorted into complete order by Country, State, City; and partial order by street and address number. This permits the high speed labelling machine to print ~~generate~~ labels from the computer file in ~~such~~ a sequence that permits the magazines to be ~~are~~ labelled in sequence and in groups according to proper batches for shipping and mailing. In order to obtain special mailing rates and privileges, the post office now insists that the publishers deliver magazines to them sorted and bundled in zip code sequence, which is roughly comparable to the sequence sort mentioned above. Consequently, most publishers ~~will probably want to~~ include the zip code in their keylines & account numbers.

The particular elements used for the keyline vary with each publisher, depending on such things as the size of the file and type of computer equipment

available. One publisher's keyline ~~other than~~ different from that shown in Fig. 3-7, constructs a keyline from the following elements:

all five digits of the zip code
three digits of the street address
three characters in the street name
first initial of the subscriber
first three letters of the subscriber's last name

Now that you know how it is done, try to decipher the mailing labels on the next few magazines you receive, in order to see what elements were used.

Some account number schemes will incorporate some personal information about the individual. One particular scheme reportedly used for driver registration files at several motor vehicle bureaus uses elements of the name, as well as date of birth, sex, eye color, and height to construct a code number with less than 0.025% duplication.* (*IBM Corp., "A Unique Computable Name Code for Alphabetic Account Numbering," Brochure F20-8052-1 of the IBM Corp., Data Processing Div., White Plains, N.Y. (1966)). One useful feature of this derived code number is that it is about 90% accurate in sorting the entire file into alphabetic sequence when sorting only on the code number. This system takes advantage of some statistical information known about last names used by people in the United States, namely that about

50% of the people share about 1500 most common names. Using these names, a table has been constructed to divide the name list into equal intervals based on the second through sixth letters of the last name, with a four digit number assigned to each interval. Similar tables are also used for first names and middle initials. An example of this code is shown ~~in~~ below.

~~Fig.~~ ___.

5] Harry ~~Henry~~ L. Sankin, born Dec. 12, 1918, ~~having~~ brown eyes, and 5' 8" tall. = S0437 31673 ~~121218~~ £1825

where:

S = First letter of last name

0437 = code number corresponding to ANKIN in equal interval last name table

316 = code number corresponding to HARR in equal interval first name table

73 = code number corresponding to L in middle initial table

d = month of birth (Dec.)

18 = year of birth (1918)

2 = male - brown eyes

5 = height range (~~5'8"~~ 67-69 inches)

9. Biographic Files Coles

Many organizations have a problem of monitoring files or records for names of individual persons (e.g., membership lists, author indexes, criminal records, genealogical records, telephone directories, retail credit records) in situations where it is cumbersome or impossible to use a more systematic identification system. Name files present many difficult problems for the file coders and users. A few of the problems are:

1. Some names may be used as either a first name or surname, or as part of a title:

   John Charles Thomas, vs Thomas Kent

   John King, vs King John

   Dean Rusk

2. Compound names:

   Garcia-Lopez

   Mary Sidney-Smith

3. Particles and Titles:

   Bill de Funiak II

   Von Bismark, Otto

   N. S. D'Andrea DuBois, Jr.

all from #5 →

4. Portions of the full (or current) names, ~~may be~~ unknown.

Fidel Castro vs
Patti Page vs

3. Some of the computerized airline reservation systems currently keep the passenger's name as part of the seat reservation (record). These particular name files are extremely susceptible to spelling errors introduced by such things as not too careful handling by secretaries, travel agents, and reservation clerks; and by noise in the communication links. Because these systems have higher error rates than more carefully controlled biographic files, special programs have been incorporated to find the "next closest" name or names, in the event that the record for the original name can not be found.

5
6. Variations in transliteration from a foreign language:

(add to list in 3.)

De-Sota-y-Rodiquez
Mario Salas, Ing.

There does not appear to be any systematic way to examine a part of a name and be able to unambiguously identify it as the surname or family name. However, some studies have been made regarding computer techniques to automatically identify personal names in text material* (*ref. C. G. Borkowski, "An Experimental System for Automatic Identification of Personal Names and Personal Titles in Newspaper Texts," AMDO, Vol. 18, No. 3, (July 1967)). fn 1/6

"A System for Automatic Recognition of Names of Persons in Newspaper Texts," Report RC-1563 of the IBM Watson Research Center, Yorktown Heights, N.Y. (3 March 1966). The problem is further complicated by the fact that the same person may use several variant forms for the same name. For example,

S. Dunn
S. Ann Dunn
Sharry Dunn
Sharron Dunn

This makes it difficult to rely on using only a straight alphabetic file.

The problem of handling written names is less complicated than the problem of handling the same collection of spoken names. Here the problem is further complicated by the words that may sound the same, for example:

Smith	Brown
Smithe	Browne
Smythe	Browen

One common approach here is to file the records under a phonetically coded version of the name. Filing by a phonetic version of the name has the effect of grouping together the records with sound alike names. One widely used method of phonetic coding is the Soundex system.

~~Soundex Code~~. The conventional Soundex system for manual files converts a name to a code word having one alphabetic character and three numeric characters.[12] It is possible to generate larger code

[12] Remington Rand, *Soundex—Foolproof Filing System for Finding Any Name in the File*, Brochure LBV809 (undated). Remington Rand, *Idem Sonans Says It's Legal (Soundex)*, Brochure LBV528 (undated).

words, but this is seldom done. The conversion rules are these:

1. Always retain the initial letter of each name.
2. Drop out A, E, I, O, U, Y, W, and H.
3. Assign the following numbers to the remaining similar-sounding sets of letters:

   B, F, P, V = 1
   C, G, J, K, Q, S, X, Z = 2
   D, T = 3
   L = 4
   M, N = 5
   R = 6
   Insufficient consonants = 0
   (e.g., DARLINGTON = D645)

4. Special cases:
   i. If there are insufficient letters, fill out with zeros (e.g., MORAN = M650).
   ii. Drop out the second letter in a letter pair (e.g., KELLEY = K400).
   iii. Drop out adjacent equivalent letters (e.g., JACKSON = J250).
   iv. Drop out adjacent equivalents of the first letter (e.g., LLOYD = L300).

Most of the methods of shorthand note taking utilize such phonetic coding (without numbers) to reduce the bulk of material to be written. These shorthand methods obtain further condensation by using a family of special graphic symbols to represent the phonetics.

The use of phonetic code has the disadvantage of destroying the alphabetical sequence of the file, as well as increasing the size of the subsection to be searched for a particular entry.* (*ref. Hewes, W. L. and K. H. Stow,

DAPR,

"Information Retrieval by Proper Name," Vol. 7, No. 6, pp. 18-22 (June 1965)),
the phonetic code has been used primarily for coding of personal names, however in at least one case it has also been used to index a collection of drug information. & (& Bolt, R.A. et. al., "Some Information Indexing Techniques in a Real-Time Hospital Computer System", JCHD, Vol. 6, no. 1, pp. 57-60 (February 1966).

One alternative to the phonetic file is the straight alphabetic file, with cross-references included to suggest other spellings to be checked for the same sounding name. Such a spelling variation cross reference structure is sometimes built into telephone directories.

Another possibliity is to record and search on some additional information elements in addition to the name. For example, in consulting a telephone book to find a friend, J. Smith, we might find that there are three people of the same name in the directory, and that we finally make the identification on the basis of additional elements of information, (e.g., the Smith on 14th Street). This technique is often used, with different elements of information chosen for specific applications. For example, some of the tags (in addition to the name) that might be used for further identification in a name file are:

name variant or alias	birth date
address	birth place
citizenship or nationality	college attended and degrees granted
organization or profession	physical features (e.g., color of eyes, sex)
titles (e.g., Dr., Miss)	
organization name	

In one application for name codes at a motor vehicle bureau, four elements of information were used in addition to the name (date of birth, sex, eye color, and height) and typically resulted in name codes with less than 0.025 percent duplication. *(* "a Unique Computable Name Code for Alphabetic Account Numbering." Report F20-8052-1 of the IBM Corp., Data Processing Div., White Plains, N.Y. 1966. 27 pp.)

The biographic elements of descriptive information found useful in the computer matching of large biographic files by one investigator concerned with matching death certificate records* (*ref. N. S. D'Andrea DuBois, Jr., "A Document Linkage Program for Digital Computers," BEHS Vol. 10, No. 3, pp. 312-319 (July 1965)) were:

Last name	day of birth	middle initial of spouse
First name	year of birth	middle name of mother
Middle initial	place of birth	social security number
Month of birth	first initial of spouse	sex

These elements are matched for corresponding records from two different files in order to find those records with a high degree of coincidence.

The elements of information felt to be sufficient for biographic identification by the Genealogical Society of the Church of Jesus Christ of Latter-Day Saints are:

Name	geographic location of individual
Date of birth	relationship to others in the family
Date of death	

Their file now consists of several million biographic records, with one of the largest microfilm storage systems in operation today.

as mentioned earlier,

Phonetic systems tend to put names together that do not logically belong together. Furthermore, depending on the way the file was built, you may not have a full and complete reference listing of all the names that are in the collection.

Another non-phonetic approach is to analyze the entire collection of names in the file, and empirically assemble the existing names into groups or families, with each group containing the possible variants for a given name. For example:

Group 3220	Smith
	Smithe
	Smythe
Group 7424	Jahnsen
	Jahnson
	Jansen
	Janson
	Johnsen
	Johnson
	Jonson

Besides providing an identification number for each group of names, it might also be helpful to assign identification numbers for each variant within a group. Proponents of this approach have suggested that once the

name groupings have been successfully established by a linguist-analyst, it should be possible, with clerical labor, to achieve better search results than might otherwise be achieved with a phonetic system.

SPECIAL CODING

Section no. assigned to this point

E. Special Coding Methods

1. Check Digits

In many file maintenance or data processing applications, it is extremely important that some types of information be error-free. A bank would be very concerned, for example, if a request for a withdrawal from one account number was transformed by some keypunch or machine error into a request for a withdrawal from another account number. This could happen by a very slight error in recording or handling the account number. Transposed characters, or a wrong character in the account number could lead to considerable aggravation for all parties involved. In order to help avoid some of these situations, methods of automatic checking have been developed which will show when an error has been made. Because some extra effort is involved in using the checking schemes, they are usually used only for the data that is most critical for a particular application (e.g., account numbers).

Self-checking numbers are usually pre-calculated and appended to the basic number for the purpose of catching keypunch or transmission errors. In most cases, a single check digit is computed and applied to the original number, which can be of any length. The check digit is derived from the basic number to be checked, usually by rules programmed for the data processing equipment; or by special equipment included as special attachments to the keypunch units. The check digit is based on the value and sometimes the position of the digits in the checked number, and can be computed for numbers of any length. It is usually computed at the time that the basic number is first entered into the file, and often at any future time when there is an information transfer or transaction involving that basic number. Two common methods of computing check digits are the modulus 10 and modulus 11 methods described below. Each of these methods is used with several different modifications.

The code verification at some later time is done by recalculating the check digit and comparing that result with the digit presently in the record. A transposition or error in the recorded characters will usually result in a different check digit. This type of check is reported to catch over 90 percent of transposition and substitution errors. Two common methods of computing check digits are the Modulus 10 and Modulus 11 methods described below. Each of these methods is used with variations on a basic theme.

Modulus 10 Method

The Modulus 10 method consists of the following basic steps: (1) multiply each digit in the number by another specified weighting digit; (2) form a sum of the results of these cross-products; (3) save the units digit of the resulting sum (i.e., "cast out 10s"). Within these basic steps a number of variations have been practiced with both the weighting digits and the method of forming the check sum.

A weighting series of alternating ones and twos, with the two applied to the units position, appears to be the most common weighting ~~values~~ method. This is used in a special attachment on some IBM keypunch equipment to automatically compute and punch a check digit; * (* "IBM 24-26 Card Punch Bulletin. Self-Checking Number Feature." Report G 24-1057-0 of the IBM Corp., Data Processing Div., White Plains, N.Y. 1958; "Introduction to the IBM System/360 Direct Access Storage Devices and Organization Methods. Student Text." Report C20-1649-1 of the IBM Corp., Data Processing Division., N.Y. 1966. Chapt. 9 describes modulus 10 and modulus 11 systems.), and is used by Dun and Bradstreet in their DUNS number for their magnetic tape files.* (*Dun and Bradstreet Co.). Chemical Abstracts Service uses an increasing series of prime numbers (1,2,3,5,7,11) to compute a check digit on the number in their machine files.

In some instances the check sum is derived by a simple sum of the individual products of the multiplications of each digit with its weighting digit. The most frequent approach is to sum each of the individual digits resulting from the multiplications of the digits with their weighting digit.

For its Registry Number,

Chemical Abstracts Service uses a different weighting series (~~an~~ ascending series starting with the value equal to one in the units position) and uses the resulting units digits of the sums of the cross products as the check digit. An example of this method is given below, using 95216 as the starting number:

9	5	2	1	6
×5	×4	×3	×2	×1
45	20	6	2	6

45 + 20 + 6 + 2 + 6 = 79

and 9 is the check digit.

An example of the modulus 10 scheme in the form in which it is most often computed is given below, using 61248 as the starting number:

6	1	2	4	8
x2	x1	x2	x1	x2
12	1	4	4	16

1+2+1+4+4+1*6 = 19

Subtract this sum from the next highest multiple of 10 (hence the name modulus 10).

20-19 = 1

which is the check digit

The new number then becomes 612481.

Other examples:

45626 becomes 456269
30759 becomes 307595
73074 becomes 730747.

insert →

Modulus 11 Method

There are some indications that an improvement in the performance of a check digit scheme depends upon the number and type of weights used for the cross products, as well as the radix of the number system used in the computation.* (*Sisson, Roger L., "An Improved Decimal Redundancy Check," CACM, Vol. 1, No. 5., pp. 10-12 (May 1958); Canning, R. G., Electronic Data Processing for Business and Industry, Wiley, 1956, pp. 189-191).

This suggests that the Modulus 11 method can provide some improvement in performance over the Modulus 10 method. As with the modulus 10 method, there are variations in the form in which the method is actually used.

Different weighting factors have been used. IBM uses a cyclic series of 2,3,4,5,6,7,2,3,..., with the ascending series starting in the units position, to automatically compute check digits with a special attachement to the keypunch unit.* (*"Self-Checking Number Feature, Modulus 11, and Its Associated Self-Checking Number Generator Modulus 11." Report G-24-1022-1 of the IBM Corp., Data Processing Div., White Plans, N.Y. 1960). The Book Numbering Agency in Great Britin uses weights in a continuous ascending series from 2 to 9, starting in the units position, for a check digit in the British National Book Number.* (*"standard Book Numbering," A report of the Standard Book Numbering Agency, London. 1967; Foster, F. G., "~~SBN~~ Standard Numbering in the Book Trade," a report of the Publishers Association, London, 1967)

An example of the IBM method is given below, using 943457842 as the starting number:

9	4	3	4	5	7	8	4	2
x4	x3	x2	x7	x6	x5	x4	x3	x2
36	12	6	28	30	35	32	12	4

But contrary to their modulus 10 technique of summing digits, IBM now sums the partial products.

36 +12 +6 +28 +30 +35 +32 +12 +4 = 195

Subtract this sum from the next highest multiple of 11 (hence the name modulus 11).

198-195 = 3

which is the check digit.

This method can sometimes generate a checkdigit with a value = 10. If there is an interest in restricting the check number to a single position, then it can be substituted for another single number (e.g. 0 Zero) or letter (e.g., X) when it occurrs. Changing to a letter will of course provide better performance than changing to a number that could have been computed on the basis of different digits.

check characters

A similar approach can be taken with alphabetic data. Luhn, for example, described a technique in 1958 which he called ELCO (Eliminate and Count), but which might be more appropriately called a modulus 26 method.* (*Luhn, H. P., "Superimposed Coding with the Aid of Randomizing Squares for Use in Mechanical Information Searching Systems," Chapt. 23 in Punched Cards--Their Application to Science and Industry, 2nd ed., Casey et. al., editors (Reinhold Publishing Corp., New York, 1958).) This method was proposed as a means to furnish an added degree of discrimination for methods of word abbreviation and coding. However, this same method could also be used to generate a check character. In the originally intended use, the check character was generated by the characters that were eliminated from the original word. For example, an abbreviation technique may reduce COMPUTERS and COMPUTATION to the same root, COMPUT, with a resulting loss in discrimination. Using a check letter, the original words would abbreviate to COMPU*, the letter in the asterisk position generally being different for each original word. There are several ways to generate a check letter: (1) add the numbers that represent the alphabetic position of the eliminated letters (e.g., A = 1, B = 2, . . ., Z = 26); (2) add the numbers that represent the frequency rankings of the eliminated letters (e.g., E = 1, A = 2, . . ., Q = 26 for names); and (3) count the number of eliminated letters. If the number resulting from any of these operations is larger than 26, then cast out 26 as many times as possible; the remainder (less than 26) is then used as the entry in a table of check letters. The table of check letters could be a simple listing of the alphabet. One method for deriving a check digit is shown in the following example, in which the word ABLATION is reduced to a total of four characters by simple truncation and the addition of a check letter:

1+20+9+15+14=59=26+26+7=H

Original word A B L A T I O N

Final abbreviation A B L H

The table for the eliminated letters uses A = 1, ..., Z = 26; but the table for the derivation of the check letters uses values A = 0, ..., Z = 25 since it is possible to obtain the value zero if the first number exactly equals a multiple of 26.

Chemical Abstracts Services uses a check digit or letter on their CODEN descriptions for periodical publications. As mentioned in Sect. ___, the CODEN normally consists of five alphabetic characters (e.g., JAMAA stands for the Journal of the American Medical Association). Because these CODENS are used extensively in the files and information handling system at Chemical Abstracts Service, they felt it important to use a check character so that the possibility of later errors due to the clerical and machine data handling would be sharply reduced.

The check character is computed from all of the characters in the CODEN, and is done by converting each alphabetic character to its equivalent rank number in the alphabet (e.g., A = 1, B = 2, ..., Z = 26) and then computing the number

$$N = 11X_{L_1} + 7X_{L_2} + 5X_{L_3} + 3X_{L_4} + 1X_{L_5}$$

where 1, 3, 5, 7, 11 are the first five prime numbers, and ~~X_1 is~~ the ~~lower~~ rightmost ~~order~~ letter position of the CODEN. The resulting number N is divided by ~~26~~ 34 as many times as necessary to obtain a remainder no greater than ~~26~~ 34 ("casting out ~~26~~ 34's"). The remainder then corresponds to the ~~alphabetic rank number of the~~ check letter shown in the following table. As an example, consider the previously mentioned JAMAA.

Remainder	Check character	
1 - 26	A - Z	5
27 - 33	2 - 8	
0	9	

	J	A	M	A	A	
~~Rank number:~~	10	1	13	1	1	(Rank numbers)
~~Multiply by:~~ (insert 2 blank spaces)	x11	x7	x5	x3	x1	
	110	7	65	3	1	
~~Sum partial products:~~	110 +	7 +	65 +	3 +	1	= 186

Cast out ~~26~~ 34's: 186 = ~~7 (26)~~ 5 (34) + ~~4~~ 16 = ~~P~~ P

Using the earlier table, in which 16 corresponds to the letter P, the final expression becomes JAMAAP.

~~Final expression: JAMAAP~~

Many other data checking techniques such as limit and completeness checks, and table lookups, are used for data and file checking.* (*General Information Manual: Disk Storage Concepts, Section on Accounting Controls. Report F 20-8161 of the IBM Corp., New York. ~~Undated~~ pre-1967).

Chemical Abstracts Service also computes a check letter to be used with the Chemical Abstract Number. this is done with a weighting series of prime numbers, & a table of alphabetic letter equivalents that excludes the values of zero, one, and L in order to avoid transcription errors. The following example shows this method used for computing the check letter for the Chemical Abstract Number for 661 (i.e. Vol. 66, No. 1), with the number digits placed in the rightmost positions of an eight digit field:

6	6	0	0	0	0	0	1
x29	x19	x17	x13	x11	x7	x3	x1
174	114	0	0	0	0	0	1

174 + 114 + 1 = 289 = 12(23) + 13

23s are cast out (instead 126 because 3 letters are missing from the alphabet) and the remainder is used to find a letter equivalent in an alphabetic table with the 0, 1, L removed. In this case, 13 corresponds to the letter P which is used as the check letter.

2. PRIME NUMBER CODING

For coordinate indexing systems, a workable coding scheme can be developed in which each subject heading or Uniterm is represented by a unique prime number, and a particular document represented by the arithmetic product of all the pertinent prime numbers. In this way, the subject index of a collection would consist of a list of numbers, with one number for each file item. A search is made by arithmetically dividing the search primes, or the product of the search primes, into the number that describes a particular document. The search criteria are satisfied if there is no fractional remainder after division.

As an example, consider a system which had a code dictionary such as the following, in which each descriptor was represented by a unique prime number. A prime number is only exactly divisible by itself and one (e.g., 2, 3, 5, 7).

Descriptor	Code Number (Primes)
Abstracts	2
Aperture cards	3
Artificial language	5
Automata	7
Automatic abstracting	11
Automatic indexing	13
Automatic programing	17
Bibliographies	19
Bliss classification	23
Cataloging	29
Character recognition	31
.	.
.	.
.	.

A document which was characterized by the three descriptors *Automatic indexing*, *Bliss classification*, and *Cataloging*, would be represented by the number $13 \times 23 \times 29 = 8671$. That is, this single number, 8671, would uniquely represent the joint occurrence of those three descriptors, since with the given table, there is no other way to derive the number 8671. Each item in the file would have its subject content represented by a single concise number. For searching purposes, each of these compound numbers would be divided by the code number for the search term, to find those numbers that could be divided without giving a finite remainder. For example, if our file item represented by 8671 were to be examined during a search for *Automatic indexing*, the search examination would consist of the division operation $8671 \div 13 = 667$, which produces an answer with no fractions. Aside from the three terms used for indexing (13, 23, 29) no other terms in the code dictionary can divide into 8671 without yielding a fraction remainder.

This particular scheme is awkward and cumbersome for the human user, and is not suited for manual systems. However, its numerical nature makes it amenable to mechanization by computers or other devices that can easily form the compound numbers and do the division for the search operations. Another difficulty with the system is that as the size of the dictionary or authority list grows, so does the length of the number that represents the average document. The magnitude of the final product depends upon the number of descriptors used and if prime numbers are assigned in ascending order (2, 3, 5, . . .), then the last items in the vocabulary will be represented by large numbers, as shown in the following list:[26]

The	50th prime number is	227
	100th	523
	200th	1,217
	500th	3,559
	1,000th	7,907
	5,000th	48,593
	10,000th	104,723
	15,000th	163,819
	20,000th	224,729
	six millionth	104,395,289

[26] Lehmer, D. N., *List of Prime Numbers from 1 to 10,006,721*, Publication No. 165, Carnegie Institute of Washington, D.C. (1914).

[26] Baker, C. L., and F. J. Gruenberger, The First Six Million Prime Numbers, Microcard ~~Formulation~~ Foundation, Washington, D.C. (1959)

Armerding, G. and F. Gruenberger, A Short Table of Prime Numbers. Report P-2500. Rand Corp., Santa Monica, Calif. December 1961. AD-604 672. (a list of the first 9500 prime numbers)

If the dictionary has 10,000 descriptors, then a document with 12 descriptors would require a coding field sufficiently large to hold a number slightly less than $(104{,}723)^{12}$. The magnitude of this number is approximately $1{:}8 \times 10^{60}$, which can be represented by 61 decimal digits. A procedure of assigning the smaller prime numbers to the most frequent descriptors would reduce the average code length. In any case, the formation and division of these long numbers will pose some awkward problems for computing equipment as well as for people.

The first reference to the use of prime numbers for classification and for information retrieval was made by Fairthorne in 1947.[25] The next reference was made by Buck in 1958.[26]

[25] Fairthorne, Robert A., "The Mathematics of Classification" reprinted in Towards Information Retrieval, Butterworth, London, p. 5 (1961).

[26] Buck, R. C., *Studies in Information Storage and Retrieval: I. On the Use of Gödel Indices in Coding*, MRC Technical Summary Report No. 53, University of Wisconsin, Math Research Center, Madison, Wisconsin (October 1958), AD-207 325; and "Studies in Information Storage and Retrieval: On the Use of Gödel Indices in Coding," *American Documentation*, Vol. 12, No. 3, pp. 165-171 (July 1961).

Buck made use of a mathematical coding problem that was solved earlier by Gödel, and referred to the coding as *Gödel indices*. Kokie later described an operating system for an IBM 650 computer that used prime numbers of up to 4 digits to index a file of documents.[27] There are approximately 1200 prime numbers with 4 digits or less, so that a reasonable dictionary size can be maintained for a small file of documents. Kokie's restriction to 4 descriptors per document was necessary because

[27] Kokie, J. E., "An Experimental Numerical Information Retrieval System," paper presented at the 15th National Conference of the Association for Computing Machinery, Milwaukee, Wisconsin (August 1960).

the computer could not conveniently handle the product of larger primes.

At this same time period, Cockayne and Hyde experimented with a prime number coding system for chemical compounds, using a computer searching system, and stated that they found this to be a way to achieve a savings in storage space over more conventional systems.[54] Some 25,000 organic chemical compounds were coded with prime numbers that represented their physical and other properties. This system was indexed by 208 different structural features, and used the first 208 primes. The smaller primes were allocated to the more frequently occurring features in an effort to keep the prime product to a relatively small number. For a given chemical compound, the primes corresponding to its structural features were selected and multiplied, and the resulting product was used as the code for that chemical compound. The test for the presence of a set of features consisted of dividing the code number by a factor, formed from the primes corresponding to the particular set, and testing for zero remainder. This process selected those compounds that had at least the specified features. It was found that even with the possibility of coding 17 features per chemical, the largest code number encountered, approximately 10^{22}, was very much less than the computed worst case, approximately 10^{52}.

In a later publication, Lamm suggested that prime number coding as previously described could be usefully modified into a positional scheme, in which the code for each element is a prime number selected in such a way that the digits occupying [redacted]itions have special me[redacted]

fu54

54 Cockayne, A. H., and E. Hyde, "Prime-Number Coding for Information Retrieval," ~~Computer Journal (Great Britain)~~, Vol. 3, No. 1, pp. 21-[illegible] (April 1960). Fairthorne, R. A., "Prime Number Coding for Information Retrieval" (a review of the article by Cockayne and Hyde), ~~Computer Journal (Great Britain)~~, Vol. 4, No. 1, p. 85 (April 1961).

fu55

[redacted] E., "Prime-Number Codi[redacted] Vol. 12, No. 3, pp. 172-177 [redacted]

anized—around 1947.[50] Its principal advantage is a reduction in storage space requirements for card, tape, or computer systems. Other advantages, for machine handling, are the built-in logic capabilities and the alignment of the descriptor fields. The primary disadvantages are the possibility of ambiguities, the occurrence of false drops,[57] the requirement of coding and encoding by table look-up, and most important—the lack of a single composite indexing number that is convenient for human interpretation, notation, and retention. The false drop problem is not a serious one since code design techniques allow the frequency of false drops to be designed to any desired level.

The usual coding method consists of determining, from a code dictionary, the particular pattern of marks assigned to each relevant descriptor, and then superimposing the patterns of all these descriptors to form one composite pattern to represent the document. These patterns could be represented in many forms, such as the punched holes in an edge-notched card or numbers recorded on magnetic tape. The marks all exist only in one of two possible states (i.e., a hole is either punched or unpunched, a number is either 0 or 1). This situation is referred to as *binary* coding, which is characterized by the fact that each code element can only exist in one of two

[50] Isbell, A. F., "A Practical Application of a Punched-Card System Utilizing the Superposition of Codes," paper presented at the Division of Chemical Literature of the American Chemical Society, 114th National Meeting, Portland, Oregon (September 1948).

Mooers, C., *Application of Random Codes to the Gathering of Statistical Information*, Bulletin 31, Zator Co., Cambridge, Mass. (1949) (based on M.S. thesis, MIT, January 1948).

Mooers, C., "Zatocoding Applied to Mechanical Organization of Knowledge," *American Documentation*, Vol. 2, No. 1, pp. 20–32 (January 1951).

Mooers, C., "Zatocoding and Developments in Information Retrieval," *Association of Special Libraries and Information Bureaux Proceedings*, Vol. 8, No. 1, pp. 3–22 (February 1956).

Wise, C., "Mathematical Analysis of Coding Systems," Chapt. 21 in *Punched Cards: Their Application to Science and Industry*, 2nd ed., R. S. Casey et al., editors (Reinhold Publishing Corp., New York, 1951).

[57] "False drops" are defined as items that are located in response to a search request, but that have no relevance to the search question. They occur because the resulting indexing pattern on a particular document is not unique, and can be synthesized from several different combinations of descriptor code patterns.

3. SUPERIMPOSED CODING

Superimposed coding was first developed, analyzed, and used as a coding technique for information storage and retrieval systems—manual and mech-

states. The following illustration shows how the method might be used for the encoding of a document with 4 descriptors:

Descriptor	Code Pattern
Magnetic	010000000001100000
Computer	000000000010000011
Circuits	000000010001000010
Digital	000000000010000011
Resulting superimposed code pattern:	010000010011100011

Just as in the case of prime number coding, a single composite pattern is derived (with the aid of a code dictionary) to represent the subject indexing of that file item. However, in this case it is much more difficult to transcribe, read, or otherwise handle the composite pattern. Notice that this is really a code pattern, and not a specific number, although it could be interpreted as a binary number and converted to its equivalent in some other number base. As mentioned earlier, the pattern could be a set of punches in a card, a set of notches on the edge of a card, a set of binary numbers stored on magnetic tape, or a set of binary marks stored on a piece of microfilm. The search process consists of superimposing the patterns of the search descriptors to form a composite quiz pattern, then checking each file item to see if the quiz pattern is included in the pattern of the file item. A search is satisfied if there is a corresponding mark in the file item for every marked position in the quiz pattern.

The code patterns in the code dictionary could be established on any basis. They might be so designed as to be unambiguous and uniqually decipherable. Consider the following example in which a document is encoded with four descriptors, and each descriptor has only a single mark in its code pattern:

Descriptor	Code Pattern
Magnetic	000000000001
Computer	000000001000
Circuits	100000000000
Digital	010000000000
Resulting Code pattern:	110....01001

In this case, it would be possible to look at the resulting superimposed code pattern and tell what the component descriptors were. This is not the case with the prior example, or most of the systems in operation today. Unfortunately, an extremely large code pattern is required for coding dictionaries with large numbers of entries. If a single mark is used in each code pattern, as in the above example, then a field of 200 code positions would be required to uniqually represent 200 descriptors. A large dictionary would lead to large code patterns, and this makes it cumbersome to use. If this method is used for computer searching then the individual and composite code pattern could each be easily represented as a binary number and this would present some conveniences for machine

handling. Use of the smaller binary numbers for the most frequently used descriptors numbers would result in a smaller average composite pattern size number and would simplify some of the computer processing. However, this approach is still somewhat cumbersome to implement, especially for files with large dictionaries. For that reason, most methods of applying superimposed codes use multiple marks in the coding field, which permit a greater number of coding possibilities for a given size of coding field, and permit the coding to be done with a smaller coding pattern. This is done at the expense of increased ambiguity in the composite code pattern, because there is a definite possibility of combining different sets of allowable patterns to arrive at the same composite pattern. Under the circumstances it is not possible to look at the composite pattern and tell with certainty which descriptors made up that pattern.

the reference card is notched with the codes for 7 descriptors. The code pattern for each descriptor has 4 marks (punches) in a total field of 40 positions. That is, in this example, there are 4 marks per descriptor in a field size of 40 coding sites. And instead of showing the binary pattern of marks for each descriptor (40 digits for each descriptor, consisting of 4 ones and 36 zeros), the card shows the *location numbers* of the positions on the card that are to be punched. Thus the descriptor *camera* is represented by 1, 8, 29, 34, meaning that those 4 positions are to be punched on the card. The top of this figure shows the similar action that takes place during the superimposition of the search terms. Note that the composite search pattern is "included" in the card's indexing pattern—hence the card would be selected.

Superimposed coding has also been used with a few tabulating card systems.[34] However, superimposed coding need not be restricted to these methods of implementation, and could be used to good advantage for computer searching systems or for other special equipment.[34]

The code patterns in the code dictionary are initially assigned to the descriptors on a random basis, primarily to distribute the marking uniformly over the whole coding pattern and enhance system performance. However, this is not essential to the operation of a superimposed coding system. One interesting modification of this random binary coding might be to use certain pairs of letters taken from the descriptor itself, rather than random numbers.[32]

To date, superimposed coding has been used primarily for edge-notched card systems. The complete card and coding system developed by Mooers (see Fig. 3-9) is the best example of such a system, and is referred to as Zatocoding.[35] In this figure,

[32] Gilbert, P. T., Jr., "An Optimal Punch Card Code for General Files," *American Documentation*, Vol. 9, No. 2, pp. 84–98 (April 1958).

[33] Brenner, C. W., "Experience in Setting Up and Using the Zatocoding System," Chapt. 11 in *Information Systems in Documentation*, Shera et al., editors (Interscience Publishers, New York, 1957).

Brenner, C. W., and C. N. Mooers, "A Case History of a Zatocoding Information Retrieval System," Chapt. 15 in *Punched Cards*, 2nd ed., R. S. Casey et al., editors (Reinhold Publishing Corp., New York, 1958).

Perry, J. W., "Superimposed Punching of Numerical Codes on Hand-Sorted Punch Cards," *American Documentation*, Vol. 2, No. 4, pp. 205–212 (October 1951).

[34] Schultz, C. K., "An Application of Random Codes for Literature Searching," Chapt. 10 in *Punched Cards*, 2nd ed., R. S. Casey et al., editors (Reinhold Publishing Corp., New York, 1958).

Luhn, H. P., "Superimposed Coding with the Aid of Randomizing Squares for Use in Mechanical Information Searching Systems," Chapt. 23 in *Punched Cards*, 2nd ed., R. S. Casey et al., editors (Reinhold Publishing Corp., New York, 1958).

[35] Mooers, C. N., *The Application of Simple Pattern Inclusion Selection to Large-Scale Information Retrieval Systems*, Technical Bulletin No. 131, Zator Co., Cambridge, Mass. (April 1959), AD-215 434.

As an aid to the reader who might be interested in using such a system, a collection of unique random codes for a field of 40 positions is given in Appendix . These codes were designed for a set of four marks in a field of 40 positions, a convenient choice of parameters for most manual systems. All the reader needs to do is to assign descriptor labels to each of the listed code patterns. They could also be used for a smaller field by ignoring those code patterns that included a mark higher than the maximum available in the smaller field. They could also be used with fewer marks per pattern by ignoring one or two of the given marks in the pattern, preferably chosen randomly from each pattern in order to avoid favoring any segment of the coding field. If more numbers are needed, the reader can refer to any of the several published tables of random numbers.

Many of the characteristics of superimposed codes can be described mathematically; some of these are discussed in more detail in later sections, using the following terminology:

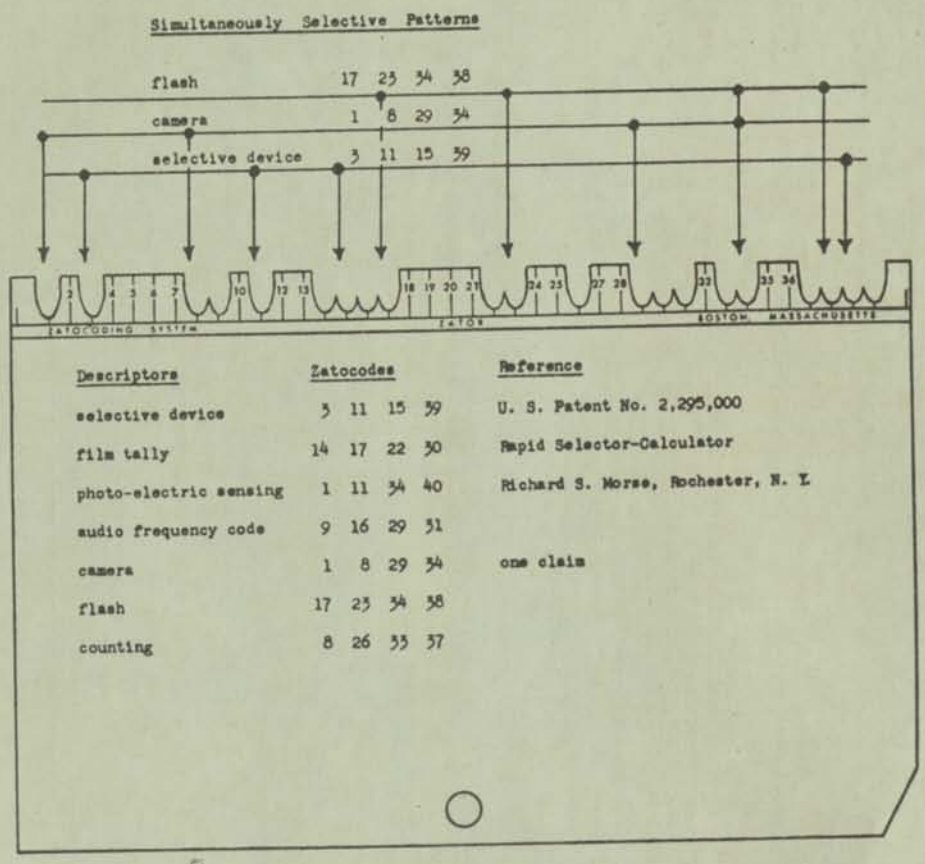


FIG. 3-9 Zatocoding, illustrated with a 5-by-8-inch edge-notched Zatocard.

	← F code positions →	
	00100001001001 . . . 001	
	. . .	d descriptors
	. . .	with m *ones* or
	. . .	marks per de-
	00001000001000 . . . 100	scriptor
	00010000000101 . . . 100	
Superimposed pattern:	00111001001101 . . . 101	M *ones* or marks in the composite field

Using this terminology, Fig. 3-9 could be characterized as having 40 code positions ($F = 40$), seven descriptors ($d = 7$) with 4 marks per descriptor ($m = 4$), and 21 marks in the composite field ($M = 21$).

Nearly all superimposed code systems represent each term or descriptor by a fixed number of marks in a coding field of fixed size. For example, edge-punched card systems normally use one full side of the card (about 40 holes or code positions) for the coding field, with a random 4-mark pattern for each descriptor. However, there are a limited number of code possibilities for m marks in a field of F positions. For a given value of F (e.g., 40 for certain edge-punched cards), the value of m should be chosen to ensure that there are at least as many different code possibilities as there are descriptors in the indexing vocabulary. The actual number (N) of different code possibilities for m marks in a single field of F positions is given by simple combinatorial analysis as the binomial coefficient, expressed below in some of its common forms.[36]

$$N = \binom{F}{m} = \frac{F!}{m!(F-m)!}$$

[36] Feller, W., *An Introduction to Probability Theory and Its Applications*, Vol. 1, 2nd ed. (John Wiley and Sons, New York, 1959).

Parzen, E., *Modern Probability Theory and Its Applications* (John Wiley and Sons, New York, 1960).

Finding the Number of Different Code Possibilities with m Marks in F Positions

where

$$F! = F(F-1)(F-2)\ldots 1$$

and

$$0! = 1$$

This means, for example, that there are 10 different ways to mark 2 holes in a field with 5 hole positions. That is,

$$N = \binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{(2\cdot 1)(3\cdot 2\cdot 1)} = 10$$

Using this formula, the number of unique code possibilities was computed for several values of m and F with the results shown in Table 3-8. When the field size is an even number, the maximum number of unique combinations is obtained when m, the number of marks per descriptor, is one-half of the field size (i.e., $m = F/2$). If the field size is an odd number, the maximum number of combinations is obtained when the number of marks per descriptor is $\frac{1}{2}(F \pm 1)$.

The actual choice of F and m is usually a compromise among several conflicting objectives. It is convenient, especially for manual systems, to use a small value of m, to simplify the marking and punching. However, a large enough value of m must be used to provide the necessary number of code combinations. It will be shown later that the values of F and m are usually chosen in a systematic way to ensure an upper bound on the number of false drops that will occur during searching.

Number of Marks in the Superimposed Field

If each descriptor uses m marks in a single field, and d descriptors are superimposed, then the total number of marks, M, that actually appear in the final pattern will lie in the range of m and md. Estimates of the mean or average number of marked positions have been made by Mooers and Wise. One estimate suggests that the average number of marked positions will be given approximately by [37]

$$M \approx F - F\left(\frac{F-1}{F}\right)^d$$

Another estimate suggests that the average number of marked positions will be given approximately by [38]

$$M \approx F(1 - e^{-md/F})$$

[37] Wise, previous citation in Casey, p. 448 (fn. 30).

[38] Mooers, C. N., *The Exact Distribution of the Number of Positions Marked in a Zatocoding Field*, Bulletin No. 73, Zator Co., Cambridge, Mass. (1952).

TABLE 3-8

Number of Unique Code Combinations (N) for m Marks in F Positions

		Number of Marks per Descriptor (m)									
		1	2	3	4	5	6	7	8	9	10
	1	1									
	2	2	1								
	3	3	3	1							
	4	4	6	4	1						
Field	5	5	10	10	5	1					
Size (F)	6	6	15	20	15	6	1				
	7	7	21	35	35	21	7	1			
	8	8	28	56	70	56	28	8	1		
	9	9	36	84	126	126	84	36	9	1	
	10	10	45	120	210	252	210	120	45	10	1
	15	15	105	455	1,365	3,003	5,005	6,435	6,435	5,005	3,003
	20	20	190	1,140	4,845	15,504	38,760	77,520	125,970	167,960	184,756
	40	40	780	9,880	91,390	658,008	—	—	—	—	—
	50	50	1,225	19,600	230,300	2,118,760	—	—	—	—	—

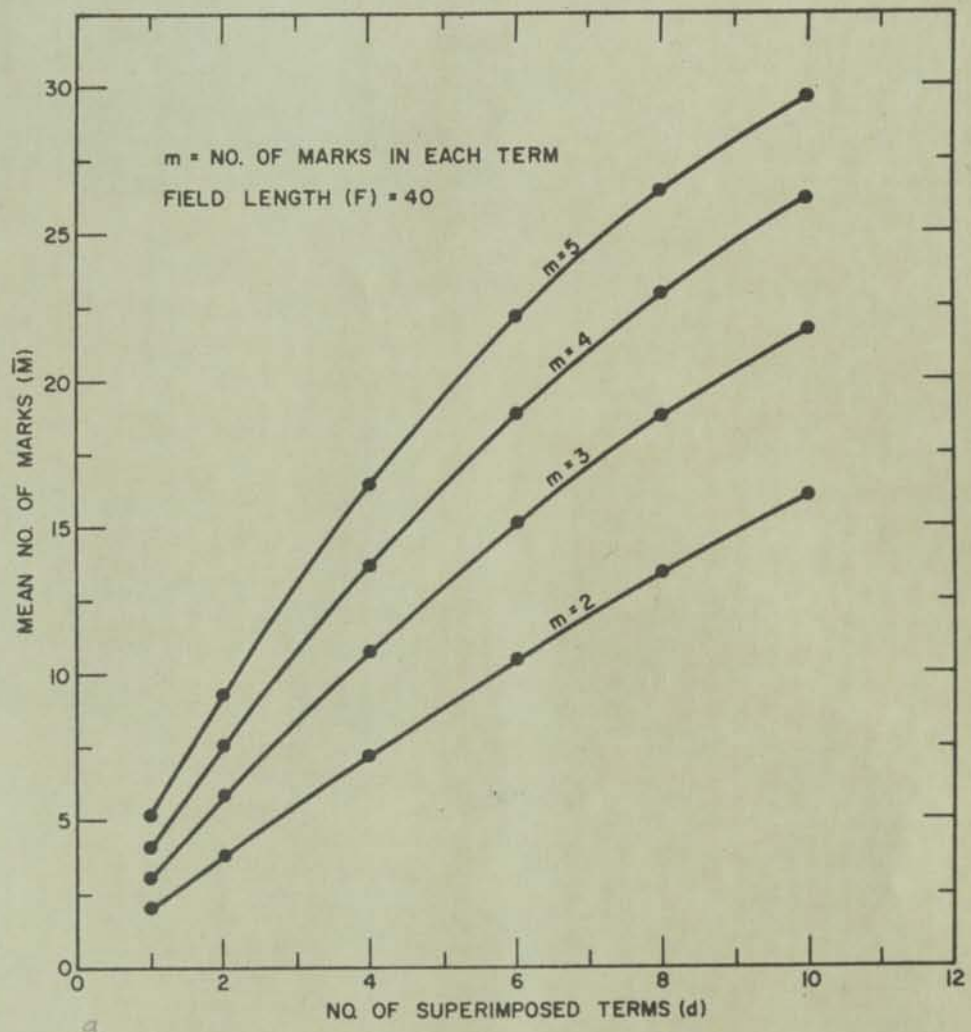


FIG. 3-10 Mean value ($\bar{M}$) of the distribution of marks in a field of 40 places resulting from the superimposition of d terms, each with m randomly placed marks.

Both of these expressions are only estimates. However, various approaches for obtaining the exact solution for the complete distribution of marked positions as well as for the mean value have been suggested and formulated explicitly.[39] There are some differences in the mathematical assumptions used by the different investigators—hence the results, and the computation procedures are somewhat different. The exact solutions are rather complicated functions, and for the sake of brevity are not repeated here. It is important to remember that these solutions are the exact solutions to a model that, unfortunately, does not exactly represent the true operational situation. To date, all of the models assume random, uniformly used descriptors, with no correlation between descriptors. All of these models differ from the real-life situations, in which some descriptors are used more than others, and in which the descriptors are not completely independent of each other—for indexing or for searching. Although no detailed study has been made to show how well these models approximate the real situations, preliminary studies indicate that they

[39] Goldberg, J., et al., *Multiple Instantaneous Response File*, Final Report RADC-TR-61-233, Stanford Research Institute, Menlo Park, California (August 1961), AD-266 169.

Orosz, G., and L. Takács, "Some Probability Problems Concerning the Marking of Codes into the Superimposition Field," *Journal of Documentation*, Vol. 12, No. 4, pp. 231-234 (December 1956).

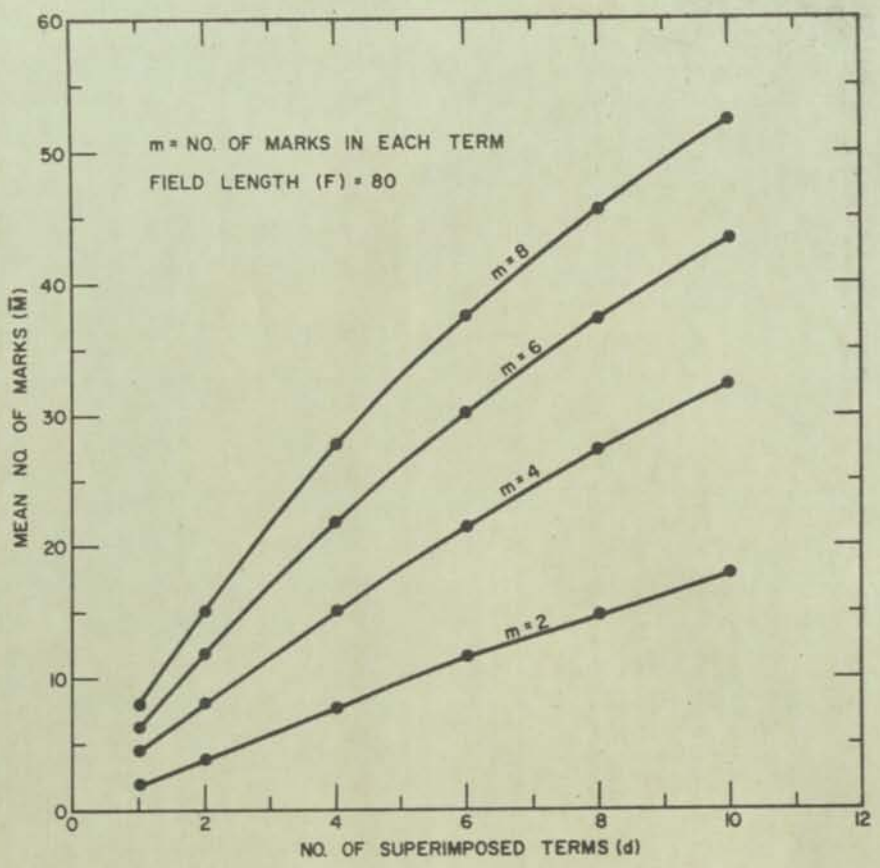


FIG. 3-11 Mean value ($\bar{M}$) of the distribution of marks in a field of 80 places resulting from the superimposition of d terms, each with m randomly placed marks.

provide a good first approximation. A few examples of the exact solutions of the mean number of marks, and their distributions in a coding field are given in Figs. 3-10, 3-11, and 3-12.[40]

Risk of False Drops

There is a finite possibility that the superimposition of several descriptors could yield a pattern that included the code pattern of one or more other descriptors. In this way, the final code pattern may appear to include irrelevant code terms. There is no way to look at the final code pattern in all instances and say with certainty exactly which descriptors were superimposed. The descriptors synthesized by the superimposition operation may result in an irrelevant response during a search. The possibility of this false drop is influenced by the coding and searching parameters, and can with proper design, be made as small as desired. The probability that a random interrogation will result in a false item being selected is often expressed as a "dropping fraction." One estimate of the upper bound to the probability of selecting an unwanted file entry is[41]

$$F_d \approx \left(\frac{M}{F}\right)^m$$

It has recently been shown that this upper bound is a good approximation which overestimates the exact random selection probabilities—thus it could be used as a conservative estimate to demonstrate the plausibility and general behavior of the code, even though it may be rather conservative for actual design.[42] The following approximation for the

[40] Goldberg report cited previously in this chapter, fn. 39.

[41] Mooers, C., *Zatocoding for Punched Cards*, Bulletin No. 30 of the Zator Co., Cambridge, Mass. (1950).

[42] Goldberg report cited earlier, fn. 39.

dropping fraction was derived from a different approach from that used for the previous approximation:[68]

$$F_d \approx \frac{\binom{M}{m}}{\binom{F}{m}} = \frac{\dfrac{M!}{m!(M-m)!}}{\dfrac{F!}{m!(F-m)!}}$$

An exact, and rather involved expression for the dropping fraction has been derived by Singleton, who has also computed the fractions for various code combinations.[69] The results of these computations indicated that the Wise approximation is too optimistic—frequently by several orders of magnitude—primarily because the derivation used the mean number of marks in the field instead of the exact distribution. The reader is cautioned that computed tables of dropping fractions that have used this expression, or variations of it which consider only the mean number of marks in a field, are probably too optimistic for serious code design work.[70] The exact computations for the probability of random selection for several different field sizes are given in Figs. 3-13, 3-14, and 3-15. These calculations appear to support Mooers' and Wise's rule that for single-field coding, a code field is best used when one-half of the places are marked.

[68] Wise, C. S., "Mathematical Analysis of Coding Systems," Chapt. 21, p. 457 in *Punched Cards*, 2nd ed., R. S. Casey et al., editors (Reinhold Publishing Corp., New York, 1958).

[69] Goldberg report cited earlier in this chapter, fn. 39.

[70] Taube, M., A. Kreithan, and L. B. Heilprin, "Superimposed Coding for Data Storage with an Appendix of Dropping Fraction Tables," Chapt. 5 in *The Mechanization of Data Retrieval*, Vol. 4 of the Studies in Coordinate Indexing Series (Documentation, Inc., Washington, D.C., 1957).

Stiassny, S., "Mathematical Analysis of Various Superimposed Coding Methods," *American Documentation*, Vol. 11, No. 2, pp. 155–169 (April 1960).

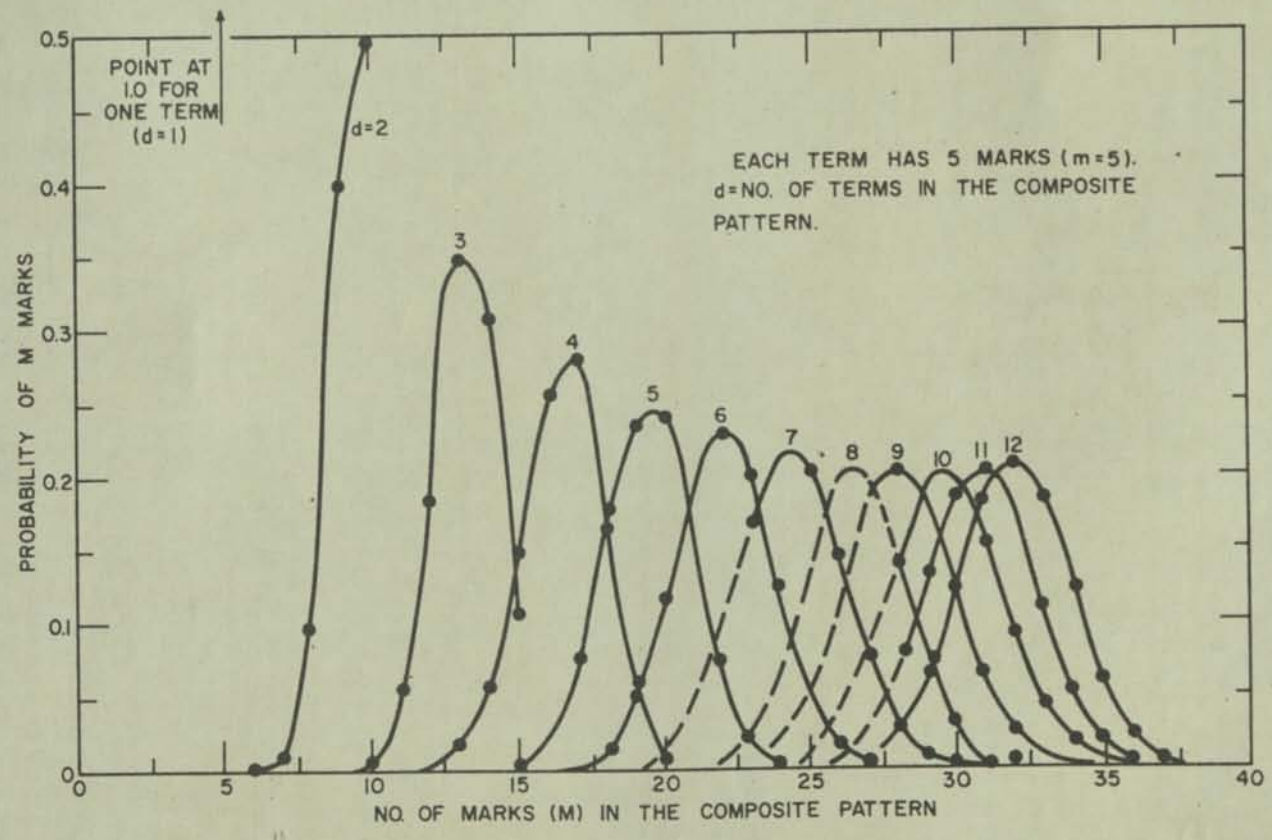


FIG. 3-12 Distribution of marks in a field of 40 places from superimpositions of *d* descriptors.

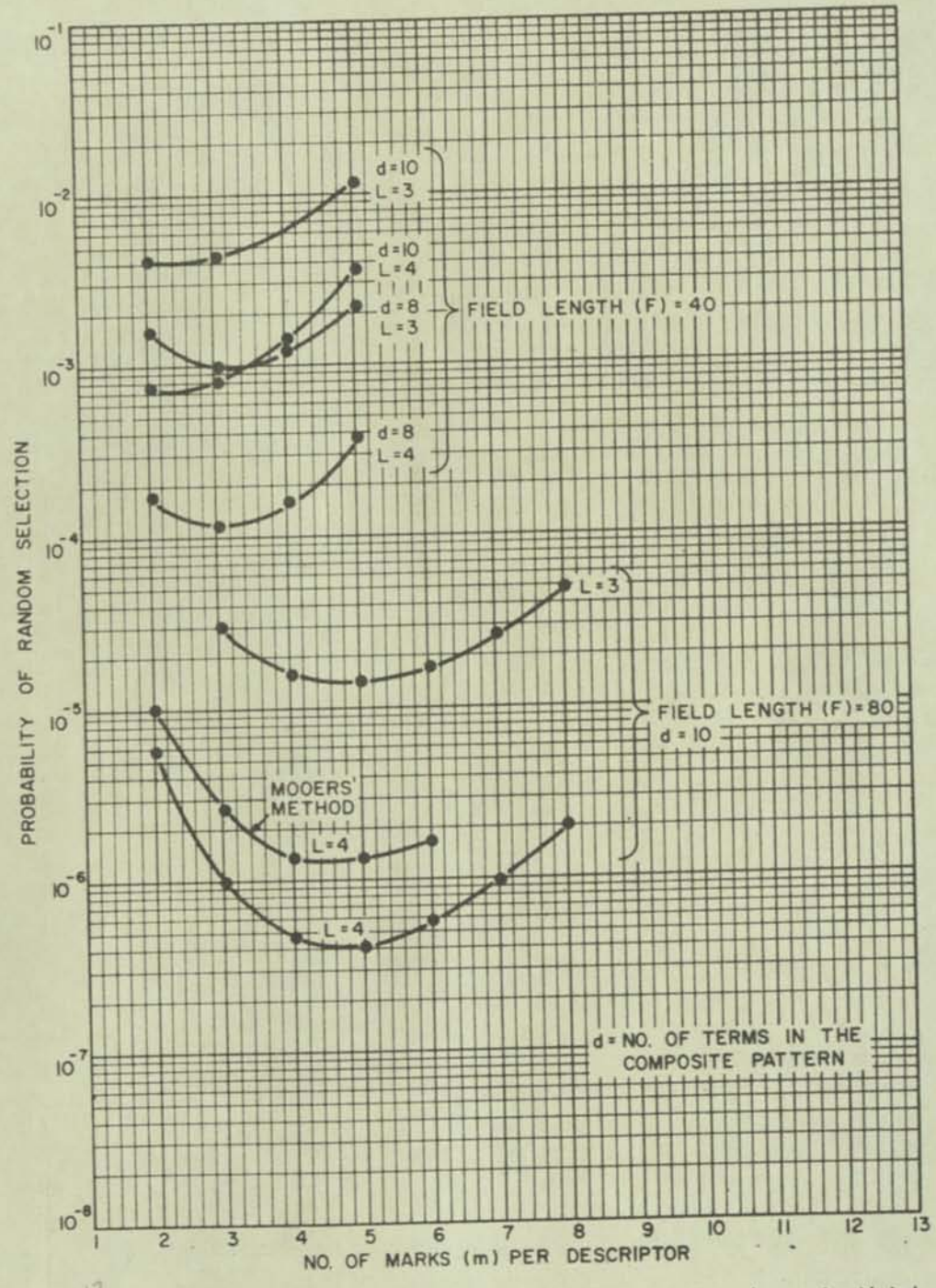


FIG. 3-13 Probability of random selection of a file item composed of d descriptors by a quiz with L descriptors, for various numbers of marks in basic descriptors.

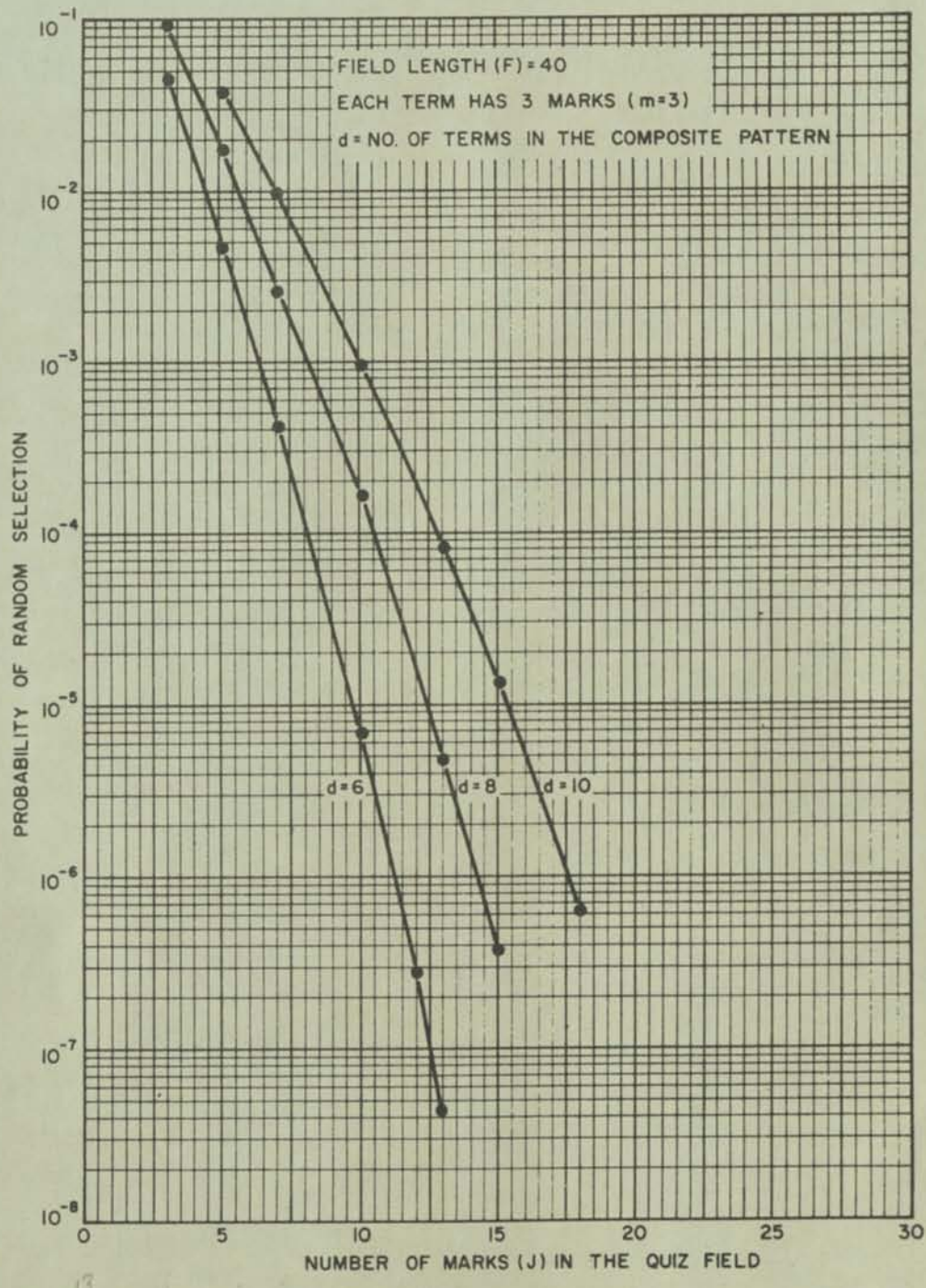


FIG. 3-14 Probability of random selection of a file item composed of *d* descriptors by a quiz with *J* marks.

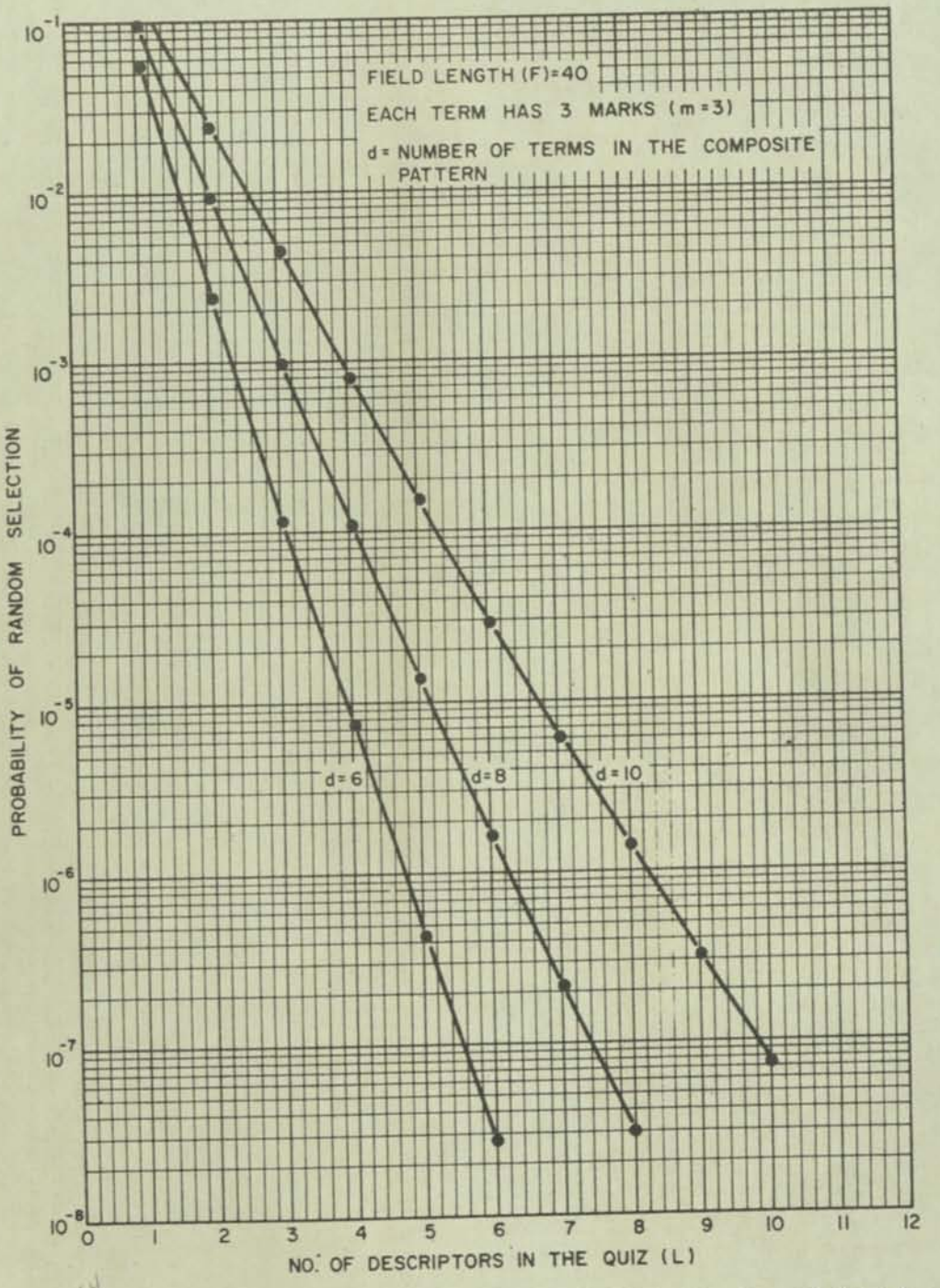


FIG. 3-15 Probability of random selection of a file item composed of d descriptors by a quiz with L descriptors.

How to Design the Superimposed Code

For relatively small retrieval systems, the user can generally adapt the systems and code parameters found to be successful by other users. However, for newer retrieval systems that require high performance of the superimposed coding system, a special study and code design may be in order. The design procedure is relatively simple, and considers the following parameters:[46]

- C the number of items in the total collection
- L the anticipated lower bound of the number of descriptors normally used for searching
- M the anticipated upper bound of the number of descriptors normally used for indexing
- R the tolerable noise ratio $= E_{max}/C$
- E_{max} the maximum number of false drops with L search descriptors
- F the length of the single fixed field for the superimposed code

In terms of these parameters, each descriptor code pattern should contain m marks (or binary ones), where

$$m = \left\langle \left(\frac{1}{L}\right)(-\log_2 R) \right\rangle$$

$$= \left\langle \left(\frac{1}{L}\right)(3.31)(-\log_{10} R) \right\rangle$$

where the symbols $\langle\ \rangle$ mean that the nearest integral value is to be taken. The least number of sites (F) that must be used to contain M descriptors is

$$F = \langle\ 1.445mM\ \rangle$$

For a sample calculation, assume the following parameters:

File size (C) = one million items

Minimum number of descriptors used for searching (L) = 3

Maximum number of descriptors used to index each item (M) = 12

Maximum number of false drops tolerable with L search descriptors (E_{max}) = 100

Tolerable noise ratio (R) = E_{max}/C

$$R = \frac{E_{max}}{C} = \frac{100}{10^6} = 10^{-4}$$

$$m = \left\langle \left(\frac{1}{L}\right)(3.31)(-\log_{10} 10^{-4}) \right\rangle$$

$$= \left\langle \frac{3.31}{3}(-1)(\log_{10} 10^{-4}) \right\rangle$$

$$= \left\langle \frac{3.31}{3}(-1)(-4) \right\rangle$$

$$= \langle\ 4.41\ \rangle$$

$$= 4$$

and

$$F = \langle\ 1.445\ (4)\ (12)\ \rangle$$

$$= \langle\ 69.36\ \rangle$$

$$= 69$$

Repeating this computation procedure for several different values of M, while keeping the same values of C, L, and E_{max} for this example, gives the following results:

Max. No. of Descriptors Used to Index Each Item (M)	Required No. of Marks per Descriptor (m)	No. of Code Positions Required (F)
3	4	14
6	4	29
12	4	69
20	4	96
40	4	191

The size of the coding field required, F, also varies with the size of the file. This slight variation is shown in Fig. 3-16, which illustrates the degree to which the specification for E_{max} influences the size of coding field required.

[46] Mooers, C. N., *The Application of Simple Pattern Inclusion Selection to Large-Scale Information Retrieval Systems*, Technical Bulletin No. 131, Zator Co., Cambridge, Mass. (April 1959), AD-215 434.

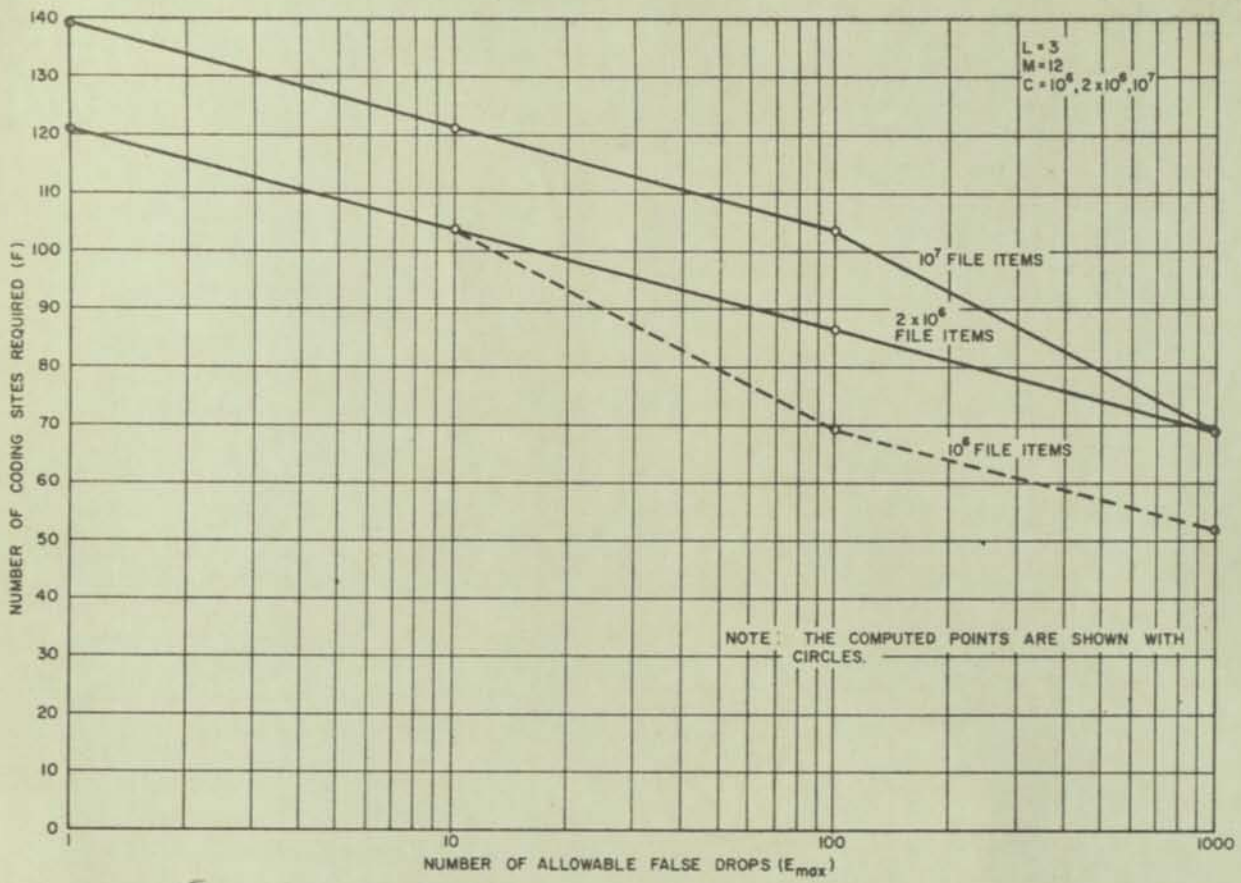


FIG. 3-16 Superimposed coding: memory requirements as a function of the number of allowable false drops.

D. Equal Interval Coding

Another example of an assigned coding scheme is the use of an empirically derived table to code surnames or subject headings into equal interval groups. Such a scheme is useful, because in many practical applications it may be desirable to break a file or list of people's names into groups of equal size. This is easy to do if the file or list already exists. However, it may be necessary to establish the division points before the file items are collected, and thus before the characteristics of the file can be known. An example would be the establishment (by the initial letters of the surname) of equal-size groups of new students to be registered during a fixed number of time periods available for class registration. These situations can be handled by using empirical data like those in Table 3-9 which show the statistical distributions of names that have been observed with other representative files. Table 3-9 is divided alphabetically into 100 groups, each group representing approximately 1/100 of the entire file. If, for example, a large and unknown group of people is to be divided into approximately equal 10 groups, then the end of every tenth group in Table 3-9 could be used as a separation point. ~~More detailed~~ breakdowns in greater detail than Table 3-9 are available for surname coding.* (*Baehne, G. W., Ed., Practical Applications of the Punched Card Method in Colleges and Universities," Columbia University Press. 1935. ~~Table~~ a code table from this book is reprinted in "10,000 Division Code for Proper Names." Report C20-8071-0 of the IBM Corp., Data Processing Div., White Plains, New York. 1961.)

A second example of the use of an equal-interval schedule is to assist in the establishment of identification or call numbers for library books to help group the books on the shelves within a particular subject classification in alphabetic order by the author's name, and to provide unique identification numbers (call numbers) for each book. Such call numbers have been derived in many libraries by using equal-interval tables such as the ones described in this section. The tables most commonly used for the establishment of library call numbers have been Cutter-Sanborn tables,

first developed by C. A. Cutter in the 1880's and based on lists of author's names, and later modified by K. E. Sanborn.* (*C. A. Cutter's Alphabetic-Order Table -- Consonants except S Altered and Fitted with Three Figures by Miss Kate E. Sanborn (Library Bureau, Boston, 1896). With the Cutter-Sanborn tables, a call number is derived by taking the first letter of the author's surname, and adding one or more decimal digits taken from the tables for that surname. Several other equal-interval tables have been developed for library call numbers.* (* W. C. Berwick Sayers, A Manual of Classification for Librarians and Bibliographers, 3rd ed., rev., Chapt. 23 (Grafton & Co., London, 1955).)

TABLE 3-A

Numerical Breakdown for Coding Names into Equal-Size Groups

Aa – Alk	01	Ell – Eu	27	Lan – Led	52	Rip – Rog	76
All – Andq	02	Ev – Feq	28	Lée – Lewi	53	Roh – Rue	77
Andr – Aul	03			Lewj – Lol	54	Ruf – Sans	78
Aun – Ball	04	Fer – Flem	29	Lom – Lym	55		
		Fleu – Fram	30	Lyn – Magm	56	Sant – Schn	79
Balm – Bar	05	Fran – Gaa	31			Scho – Sec	80
Bas – Belk	06			Magn – Marsg	57	Sed – Shel	81
Bell – Beu	07	Gab – Gel	32	Marsh – Meat	58	Shem – Simo	82
Bev – Bln	08	Gem – Gla	33	Meau – Meem	59	Simp – Smith, G.	83
Blo – Bour	09	Glb – Grac	34	Meen – Meni	60	Smith, H. – Sot	84
Bous – Breec	10	Grad – Grid	35	Menj – Mey	61	Sou – Stans	85
Breed – Brown, H.	11	Grie – Hag	36	Mez – Mir	62	Stant – Stev	86
Brown, I. – Bun	12			Mis – Mord	63	Stew – Stuf	87
Buo – Calc	13	Hah – Ham	37	More – Muk	64	Stug – Tag	88
		Han – Harrisa	38	Mul – Nea	65		
Cald – Carq	14	Harrisb – Hay	39			Tah – Thei	89
Carr – Chanc	15	Haz – Hero	40	Neb – Niel	66	Thej – Tis	90
Chand – Claq	16	Herp – His	41	Nielm – Oha	67	Tit – Tuq	91
Clar – Cofe	17	Hit – Hon	42			Tur – Vaq	92
Coff – Con	18	Hoo – Hua	43	Ohb – Paf	68		
Coo – Coy	19	Hub – Hya	44			Var – Walk	93
Coz – Culp	20	Hyb – Jam	45	Pag – Patte	69		
Culq – Dau	21			Pattf – Pes	70	Wall – Wat	94
		Jan – Johnson, R.	46	Pet – Pk	71	Wau – Wer	95
Dav – Dek	22	Johnson, S. – Kac	47	Pl – Prh	72	Wes – Wh	96
Del – Dim	23			Pri – Rand	73	Wi – Wilm	97
Din – Doy	24	Kad – Kel	48			Wiln – Wom	98
Doz – Dye	25	Kelm – King	49	Rane – Ren	74	Won – Wyc	99
Dyf – Elk	26	Kinh – Kor	50	Reo – Rio	75	Wyd – Z	100
		Kos – Lam	51				

Source: This breakdown is a modification of the results of a detailed study of 63,000 names by the Records Bureau of Stanford University. To simplify the use of this table, some of the groups have been rounded off to simpler demarcation points.

References on Linguistics & Text Analysis

Herdan, G. "Quantitative Linguistics" Butterworth. 1965.

Paisley, W. J. "The Effects of Authorship, Topic, Structure, and Time of Composition on Letter Redundancy in English Texts," Journal Of Verbal Learning and Verbal Behavior Vol. 5, No. 1, pp. 28-34, (February 1966).

Dolby, J. L. and H. L. Resnikoff. "On the Structure of Written English Words," LANG, Vol. 40, no. 2, pp. 167-196, (April-June 1964).

Alberga, C.N., " String Similarities and Misspellings," CACM, Vol. 10, No. 5, pp. 302-313 (May 1967).

Barnard, G.A., "Statistical Calculation of Word Entropies for Four Western Languages," IREG, Vol. 1, No. 1, pp. 49-53 (March 1955).

Blair, C. R., "A Program for Correcting Spelling Errors," IFCN, Vol. 3, No. 1, pp. 60-67, (March 1960).

Burton, N. G., and J. C. R. Licklider, "Long Range Constraints in the Statistical Structure of Printed English," AJPC, Vol. 68, No. 4, pp. 650-653 (December 1955).

Carson, D. H., "Letter Constraints Within Words in Printed English," KYBE, Vol. 1, No. 1, pp. 46-54 (January 1961).

Chapanis, A., "The Reconstruction of Abbreviated Printed Messages," JEPS Vol. 48, No. 6, pp. 495-510 (1954).

Davison, L. "Retrieval of Misspelled Names in an Airlines Passenger Record System," CACM, Vol. 5, No. 3, pp. 169-171 (March 1962).

Evans, M. W., C. K. McElwain, and F. Van Hoosen, "Machine Correction of Garbled English Text," Massachusetts Institute of Technology, Lincoln Laboratory Report (1960), AD-237 700.

Mandelbrot, B., "Simple Games of Strategy Occurring in Communication Through Natural Languages," IREG, Vol. PGIT-3, pp. 124-137 (March 1954).

Miller, G. A., "Some Effects of Intermittent Silence," AJPC, Vol. 70, No. 2, pp. 311-313 (June 1957).

Miller, G. A., and E. A. Friedman, "The Reconstruction of Mutilated English Texts," IFCN, Vol. 1, No. 1, pp. 38-55 (September 1957).

Miller, G. A., and E. B. Newman, "Tests of a Statistical Explanation of the Rank-Frequency Relation for Words in Written English," AJPC, Vol. 71, pp. 209-218 (March 1958).

Miller, G. A., E. B. Newman, and E. A. Friedman, "Length-Frequency Statistics for Written English," IFCN, Vol. 1, No. 4, pp. 370-389 (December 1958).

Newman, E.B., "The Pattern of Vowels and Consonants in Various Languages," AJPC, Vol. 64, No. 3, pp 369-379 (July 1951).

Newman, E.B., and L. S. Gerstman, "A New Method for Analyzing Printed English," JEPS, Vol. 44, No. 2, pp. 114-125 (August 1952).

Newman, E. B. and N. C. Waugh, "The Redundancey of Texts in Three languages," IFCN, Vol. 3, No. 2, pp. 141-153 (June 1960).

Shannon, C. E., "Prediction and Entropy of Printed English," BSTJ, Vol. 30, No. 1, pp. 50-64 (January 1951).

West, M. A., General Service List of English Words with Semantic Frequencies (Longmans, Green, New York, 1953).

Yngve, V. H., "Gap Analysis and Syntax," IREG, Vol. IT-2, No. 3, pp. 106-112 (September 1956).

Reference on Coding & Notation

Modern Coding Methods. Report X21-3792-6 of the IBM Corp. New York, N.Y. undated.

Barrett, J.A. and M. Grems, "Abbreviating Words Systematically," CACM, Vol. 3, No. 5, pp. 323-324 (May 1960).

Bremer, R. W., "Do It By the Numbers - Digital Shorthand," CACM, Vol. 3, No. 10, pp. 530-536 (October 1960).

Foster, M. G., "A Method of Representation, Storage and Retrieval of 13 Random Codes in a 4-Digit Number or 16 Random Codes in a 5-digit Number," CACM, Vol. 5, No. 3, p. 165 (March 1962).

Gardfield, E., "Information Theory and Other Quantitative Factors in Code Design for Document Card Systems," JCHD, Vol. I, No. 1, pp 70-75 (January 1961).

Korolev, L. N., "Coding and Code Compression," JACM, Vol. 5 No. 4, pp. 328-330 (October 1958).

Mooers, C. N., "Improvements in or Relating to Encoding Information on Punched Record Cards and Other Equivalent Media," British Patent No. 681,902 issued 29 October 1952.

Mooers, C. N. "Manipulating Intelligence with a Random Code System," Canadian Patent No. 534,926, issued 25 December 1956.

Richmond, P. A., "A Simple Mnemonic Large-Base Number System for Notational Purposes," JDOC, Vol. 14, No. 4, pp. 208-211 (December 1958).

McCutchen, C.W., "Random Code Numbers for Universal Identification of Documents," AMDO, Vol. 16, No. 2, pp. 91-96 (April 1965).

IBM Corp., Data Processing Div., White Plains, N.Y. "Raw Material Coding," Report J20-8018-0. 1959. 8 pp.

Uhlmann, W., "The Application of Random Superimposed Coding and Chain Spelling to Peek-A-Boo Cards," AMDO, Vol. 15, No. 2, pp. 89-92 (April 1964).

Reference on Chemical Coding and Notation

Ballard, D. L., and F. Neeland, "A Computer Technique for the Retrieval of Related Chemical Structures Utilizing a Special Topological Cipher," JCHD, Vol. 3, No. 4, pp. 196-201 (October 1963).

Barnard, A. J., et. al., "Computer-Orineted Chemical Names," JCHD, Vol. 6, No. 1, pp. 48-57 (February 1966).

Barnard, A. J., "Retrieval of Organic Structures from Small-to-Medium Sized Collections," JCHD, Vol. 6, No. 1, pp. 41-48 (February 1966).

Bonnett, H. T. and D.W. Colhoun, "Application of a Line Formula Notation in an Index of Chemical Structures," JCHD, Vol. 2, No. 1, pp. 2-6 (January 1962).

Bouman, H., "Computer Program for the LINCO System," JCHD, Vol. 5, No. 1, pp. 14-24 (February 1965).

Bouman, H., "Linearly Organized Chemical Code for Use in Computer Systems (Locus)," JCHD, Vol. 3, No. 2, pp. 92-96 (April 1963).

Brasie, W.C. and D.W. Liou, "Estimating Physical Properties: Chemical Structure Coding," Chemical Engineering Progress, Vol. 61, No. 5, pp. 102-108 (May 1965); and No. 7, p. 16 (July 1965).

Crane, E.M., and M.M. Berry, "Which Notation?", CENE, Vol. 33, No. 2842 (1955).

Davis, C.H., "A Simple Code for Improving the Retrieval of Information Associated with Keto-Enol Tautomers," JCHD, Vol. 6, No. 4, pp. 199-205 (November 1966).

Davison, W.H.T., and M. Gordon, "Sorting for Chemical Groups Using Gordon-Kendall-Davison Ciphers," AMDO, Vol. 8, No. ? , pp. 202-210 (1957).

Dreffin, W.M., "Chemical Coding for Information Retrieval," JCHD, Vol. 1, No. 3, pp. 44-46 (1961).

Dyson, G.M. et. al. , "Mechanical Encipherment of Chemical Ring Structures from the Random Matrix," PADO, 1963. Pt. I., H.P. Luhn, Ed., pp. 79-82 (American Documentation Institute, Washington, D.C., 1963).

Feeman, J.F., "A Novel Organizational Code for Organic Structures Based on Functional Groups," JCHD, Vol. 6, No. 3, pp. 184 -187 (August 1966).

Feldman, A. et. al., "The Automatic Encoding of Chemical Structures," JCHD, Vol. 3, No. 4, pp. 187-189 (October 1963).

Frome, J., and P.T. O'Day, "A General Chemical Compound Code Sheet Format," JCHD, Vol. 4, No. 1, pp. 33-42 (January 1964).

Frome, J., "Generic Mechanized Search System," JCHD, Vol. 2, No. 1, pp. 15-18 (January 1962).

Frome, J., "Mechanized Searching of Phosphorous Compounds," JCHD, Vol. 1, No. 1, pp. 76-90 (January 1961).

Frome, J., "Searching Chemical Structures," JCHD, Vol. 4, No. 1, pp. 43-45 (January 1964).

Geer, H.A., et. al., "The Parke-Davis Code for Chemical Structures," JCHD, Vol. 2, No. 2, pp. 110-113 (April 1962).

Geer, H.A., and C.C. Howard, "The Role of a Central Chemical-Biological Records Center in a Pharmaceutical Company," JCHD, Vol. 2, No. 1, pp. 51-53 (January 1962).

Gelberg, A. et. al., "A Program for Retrieval of Organic Structure Information, " JCHD, Vol. 2, No. 1. pp. 7-11 (January 1962).

Hiz, H., "A Linearization of Chemical Graphs," JCHD, Vol. 4, No. 3, pp. 173-180 (July 1964).

Ihndris, R.W., "Chemical Structure Fragmentation for Use in a Coordinate Index Retrieval System," JCHD, Vol. 4, No. 4, pp. 274-278 (October 1964).

Lauer, C.M., and F.R. Whaley, "A Formula Index for Silicon Chemicals," JCHD, Vol. 3, No. 3, pp. 150-153 (July 1963).

McDonnell, P.M. and R.F. Pasternack, "A Line-Formula Notation System for Coordination Compounds," JCHD, Vol. 5, No. 1, pp. 56-60 (February 1965).

Sher, ~~J. I.~~ Irving H., et. al., "Rotadox - A New Index for Generic Searching of Chemical Compounds," JCHD, Vol. 4, No. 1, pp. 49-53 (January 1964).

Silk, J.A., "An Improved System for Enumeration and Description of Ring Systems, " JCHD, Vol. 1, No. 3, pp. 58-62 (1961).

Silk, J. A., "A Linear Notation for Organic Compounds," JCHD, Vol. 3, No. 4, pp. 189-195 (October 1963).

Skolnits, H. and A. Clow, "A Notation System for Indexing Pesticides," JCHD, Vol. 4, No. 4, pp. 221-227 (October 1964).

Starker, L.N., and J.A. Cordero, "The Cyanamid Organic Structure Code and Search System," JCHD, Vol. 2, No. 1, pp. 12-15 (January 1962).

Waldo, W.H., "Searching Two-Dimensional Structures by Computer," JCHD, Vol. 2, No. 1, pp. 1-2 (January 1962).

Willard, J.R. and E.J. Malkievich, "A Working System for Retrieval of Chemical Structures, Adaptable to Pesticidal Screening Data," JCHD, Vol. 4, No. 4, pp. 211-217 (October 1964).

Rose S, 2976
Ross FR, 3213
Rosslimo OK, 2991
Rowbury RJ, 3223
Rowsell HC, 3085
Roy J, 3114
Roychoudhury R, 3311
Rozanova LS, 3143
Rozumna NM, 3170
Rubin R, 3195
Ruzsa P, 2983
Saez-Poulain S, 2987
Saffiotti U, 3359
Sahashi Y, 3099
Sakai K, 3299
Sakurai H, 3309
Salama AM, 3119
Sallis JD, 3210
Samie MA, 2986
Sandmeier M, 3319
Sansone M, 2957
Santelmann PW, 3377,3379
Sanwal BD, 3039
Saprin AN, 3145
Sarkisova MM, 3141
Sato N, 3262
Sattur PB, 2939
Saurbrey J, 3003
Schafer JF, 3325
Scherphof GL, 3019
Schischke B, 2943
Schmidt B, 3326
Schmidt CH, 3384
Schmitt H, 2968
Schnaider J, 3146
Schoenenberger GA, 3041
Schreiber K, 3166
Schryver HF, 3149
Schultz EJ, 3273
Schulz KR, 2942
Schwab JH, 3341
Sebanek J, 3165
Seligmann M, 3270
Sembdner G, 3166
Semenuk G, 3284
Sen NP, 3118
Sen SP, 3311
Sercarz E, 3367
Serchi G, 3097
Seybold D, 3264
Shah NS, 3139
Shalyapina VG, 3304
Shan'ko VM, 3179
Shanaman J, 2961
Shannon PVR, 3261
Sharpless SK, 3255
Shata MK, 2986
Shats VYa, 3370
Shepherd D, 3068
Shibata S, 3343
Shifrin MA, 3352
Shigematsu A, 3366
Shimono H, 3207
Shiobara Y, 3262
Shiro T, 2937
Shister HE, 2969
Shnaidman LO, 3285
Shorin VA, 2991
Shubik P, 3359
Shushevich SI, 3362
Siegel JH, 3100
Siekevitz P, 3211
Sillibourne JM, 3054
Silver MD, 2979
Simonyan NA, 3140
Singh VN, 3151
Sinnhuber RO, 3354
Siro-Brigiani G, 3098
Skrentny RF, 2942
Skrivanova J, 3117
Skyring A, 2977
Slopek S, 2970
Sloviter HA, 3036
Sluyser M, 3042
Smart JV, 3294
Smith JN, 3066
Smith OE, 3306
Smith RH, 3286
Smith S, 3033
Snell EE, 3057,3058
Snipe RM, 3046
So BT, 3384
Solti F, 2983
Sonnenblick EH, 3100
Soyka LF, 3356
Spitznagel JK, 3237,3341
Spolitis A, 3281
Spriggs TLB, 3094
Stackhouse HL, 3107
Stanislavskaya MS, 2991
Stanley RG, 3296
Starka L, 3030
Stavorski J, 3249
Stehbens WE, 2979
Stein AA, 2978
Stein O, 3035
Stein WD, 3071
Stein Y, 3035
Steinbrunn W, 3387
Steinhardt C, 3271
Stephens RJ, 3380
Stevens DF, 3148
Stille W, 3266
Stoehr H, 3324
Stone CA, 3249
Stout FM, 3354
Straume O, 3281
Struijk CB, 3032
Sucaciu G, 2955
Sugino Y, 3207
Sulcova J, 3030
Sussman AS, 3233
Suzuki T, 3099,3116,3299
Symes MO, 3198
Syper L, 2974
Szepsenwol J, 3336
Szirmai EA, 3006
Taira N, 3224
Takahashi M, 2937
Takahashi Y, 3353
Takamiya A, 3309
Taljedal IB, 3287
Tange JD, 3000
Tanigawa H, 3299
Tarr HLA, 3114
Tatum EL, 3026
Taylor KW, 3269
Taymor ML, 3212
Tenhunen R, 2984
Ter-Grigorova EN, 3316
Teraoka H, 3207
Terlets'ka YT, 3361
Tetreault L, 3298
Tew RP, 3054
Thomas J, 3286
Thomas WA, 3159
Thuillier G, 3109
Thuranszky K, 2958,2959,2994
Timsit J, 3247
Todd R, 3212
Toman JEP, 2960
Tominaga S, 3224
Torchiana ML, 3249
Trettin HJ, 3395
Tronche P, 2995
Troyer WG, 3110
Tsirul'nikov VA, 3176
Tsukada M, 3225
Tsukamura M, 3220
Tsukamura S, 3220
Tsukano Y, 2942
Tugarinov VV, 3144
Turetskaya RKh, 3169
Turner CW, 3153
Turner JM, 3070
Turner P, 3294
Tyulenev VI, 3361
Tzonev I, 3005
Udaka S, 3234
Uematsu T, 3125
Ungar F, 3152
Ushakova MD, 3051
Ushakova MT, 3285
Utzumi K, 3210
Van-Bergen FH, 2936
Van-Deenen LLM, 3019
Van-Der-Kloot WG, 3104
Van-Eeken CJ, 2952
Van-Haaren R, 3105
Varma DR, 3086
Vazquez D, 3222
Venkatachalam PS, 3189
Vernon-Roberts B, 3241
Vielhauer E, 2946
Vikhlyaev YuI, 3391
Villemez CL, 3328
Vivian VM, 3244
Vlassembroeck M, 2989
Vogels GD, 3016
Vohra P, 3330
Volfin P, 3038
Von-Haartmann U, 3271
Vovk SI, 3174
Vrbova G, 3256
Wacker WEC, 3041
Wagle SR, 3333
Wajchenberg BL, 3146
Waldroup PW, 3331
Wallace AG, 3110
Walton RP, 3342
Wandall HH, 3003
Wang HH, 3252
Wang SC, 3252
Ward C, 3044
Warner ED, 2999
Warren KS, 3289
Wastila WB, 3250
Weber E, 2944
Wegener K, 3267
Weisberger AS, 3289
Weisburger EK, 3274
Weisburger JH, 3274
Weiss D, 3326
Wenger HC, 3249
Wenzel M, 3300
Westhoff DD, 3105
Whitehouse MW, 3073
Whittam R, 3031
Widdowson E, 3078
Wiener M, 2978
Wiley JS, 3031
Wilkinson C, 3199
Williams KIH, 3349
Williams RH, 3283
Wilson WS, 3100
Winbury MM, 2961
Witon J, 2974
Wittorf R, 3265
Wollgiehn R, 3164
Wong ET, 3357
Wood FC Jr, 3283
Woods DD, 3223
Wright B, 3044
Yakovleva GS, 3050
Yamaguchi T, 3201
Yanishevskii FV, 2945
Yates DW, 3068
Yogi K, 3262
Yoshinari T, 3225
Yosihara H, 3350
Yotis WW, 3229
Young JR, 3219
Young LCT, 3296
Young PR, 3358
Zaheer SH, 2939
Zaidler YaI, 3167
Zalis EG, 2965,2966
Zamenhof PJ, 3329
Zamenhof S, 3329
Zeilmaker GH, 2938
Zeya HI, 3237,3341
Zhak SP, 3390
Zhukova TA, 3168
Zhukova VO, 3178
Zimmerer RP, 3308
Zlatkina AR, 3316
Zoellner N, 3265
Zotti EF, 3110

MOLECULAR FORMULA INDEX

DIGEST NUMBER	MOLECULAR FORMULA	REGISTRY NUMBER
2936	CCl_4	56235
3008		
3122-1		
3265		
2972-4	$CHCl_3$	67663
2994-2		
3067-2	CHO_2	71476
3161-1		
3208		
3368		
3387	CHO_3	71523
	CH_2O_3	
3033-3	$\cdot K$	298146
3108	$\cdot Na$	144558
2972-2	CH_4N_2O	57136
3003		
3080		
3264		
3084-2	$CH_4N_2O_2$	127071
3206-1		
3230-4		
3084-2	$CH_4N_2O_3$	686680
2994-2	CH_4O	67561
3161		
3084-2	CH_5NO	593771
3048-2	CH_5N_3O	57567
3113-6	CKN	151508
3310-1		
3114-2	CN	57125
2968	CNNa	143339
2968-3	CO_2	124389
3016		
3058-4		
3076		
3108		
3113-1		
3155-1		
3200-3		
3233-1		
3301-3		
3338-5		
3082-3	$C_2HBrClF_3$	151677
2962	C_2HCl_3O	75876
3211	$C_2HCl_3O_2$	76039
3014-3	$C_2H_2IO_2$	152341
3015-1		
3103-5		
3114-2		
	$C_2H_3ClO_2$	
3380	$\cdot Na$	3926623
	$C_2H_3FO_2$	
3224-3	$\cdot Na$	62748
3224-3	$C_2H_3IO_2$	64697
3033-5	$C_2H_3O_2$	71501
3067-5		
3070-2		
3101		
3186-1		
3220-1		
3261		
3301-3		
3338-1		
3362		
3067-5		124505
3301		
3067-5		676799
3301		
3084-2	$C_2H_4N_2O_4$	1687601
3031	C_2H_4O	75070
3216		
	$C_2H_4O_2$	
3037	$\cdot Na$	127093
3210-3		
3097-1	$C_2H_4O_3$	79210
2990-2	C_2H_5N	151564
2998-2	$C_2H_5NO_2$	56406
3017-5		
3026-1		
3070-2		
3096-1		

DIGEST NUMBER	MOLECULAR FORMULA	REGISTRY NUMBER
3107-1	$C_2H_5NO_2$	56406
3159-1		
3220		
3084-2	$C_2H_5NO_3$	584076
3329	$C_2H_5N_5O_3$	70257
3384		
3230-1	$C_2H_6N_2O_2$	3646397
2994-2	C_2H_6O	64175
3013		
3076		
3115		
3180		
3216-4		
3233-2		
3277		
3313		
3314		
3354-2		
3015	C_2H_6OS	60242
3207-1		
3385	$C_2H_6O_4S$	77781
3264	$C_2H_7NO_3S$	107357
2953-8	C_2H_7NS	60231
3047	$C_2H_8N_2$	57147
3357		
3253		107153
	$C_3H_2O_2$	
3354-1	$\cdot Na$	1977317
3033-5	$C_3H_2O_4$	156809
3261		
3221-5	$C_3H_3N_3O_2$	461892
2968-2	$C_3H_3O_3$	57603
3045-1		
3059-7		
3070-2		
3076-1		
3333		
3354	$C_3H_4O_2$	542789
3246	$C_3H_4O_3$	127173
2946		497154
3033-4	$C_3H_4O_4$	141822
3039	$C_3H_4O_6P$	73892
3081	C_3H_5NO	79061
3384	$C_3H_5N_3O_2$	3844631
2968-2	$C_3H_5O_3$	113213
3031		
3041-8		
3039-3	$C_3H_5O_4$	383868
3082-3	C_3H_6	75194
2994-2	C_3H_6O	67641
3015-4		
3185-1	$C_3H_6O_2$	79094
3185-2	$\cdot Na$	137406
3045	$C_3H_6O_3$	50215
3123		
3155-1		
3300		
3383-1		
3103-2		56826
3300-2		453178
3300		497096
3170	$\cdot Na$	72173
3076-4	$C_3H_6O_7P$	74704
3081	C_3H_7NO	79050
3070		298088
2958	$C_3H_7NO_2$	51796
2962		
3084		
2984-8		56417
2998-2		
3046-1		
3067-6		
3057-1		107959
2951	$C_3H_7NO_2S$	52904
3018-2		
3060		
3157-1		
3172-2		

DIGEST NUMBER	MOLECULAR FORMULA	REGISTRY NUMBER
3207-1	$C_3H_7NO_2S$	52904
3208-2	$C_3H_7NO_3$	56451
3368		
3084		589413
3016	$C_3H_7N_3O_3$	3844697
3123	$C_3H_7O_6P$	57045
3076-1		142109
3230-1	$C_3H_8N_2O_2$	3833509
3230-1		3833510
3230-1		3861038
3036	$C_3H_8O_3$	56815
3037		
3059-7		
3217-2		
3232-5		
3332-1		
3178-2	$C_3H_8O_3S_3$	74613
3070-5	C_3H_9NO	78966
3070-1		1674562
2953	$C_3H_9N_3S$	1190745
3350-2	$\cdot {}^1/_2H_2SO_4$	3979008
2985	$C_4Cl_2F_8$	355248
3038	$C_4F_2O_5$	3862326
3038	C_4HFO_5	3791728
2985	$C_4H_2F_8$	377366
3269-2	$C_4H_2N_2O_4$	50715
3299		
3038	$C_4H_2O_5$	149633
3301-4		
3356-3	$C_4H_3AuO_4S$	3845054
3002	$C_4H_3FN_2O_2$	51218
3221-5		
3231		
3306-7		
3335		
3221-5	$C_4H_4N_2OS$	156821
3284		
3020	$C_4H_4N_2O_2$	66228
3116		
3163		123331
3231-1		
3236-1		
3238		
3239-3		
3292-1		
3221-5	$C_4H_4N_6O$	134587
3012-1	$C_4H_4O_4$	56144
3060		
3158		
3301-2		
3286	$\cdot xC_3H_6N_2$	1119284
3059-7	$C_4H_4O_5$	149611
3113-3		
3301-2		
3058-7		305646
3058-7		2375022
3113-5	$C_4H_5O_4$	3938850
2990-1	$C_4H_6Br_3NO$	3795957
3151-4	$C_4H_6N_2S$	60560
3280-1		
3137-2	$C_4H_6N_4O_3S_2$	59665
3195-1		
3216-2	$C_4H_6O_2$	431038
3033-4	$C_4H_6O_4$	110156
3108-3		
3246-2		
3210	$\cdot 2Na$	150903
3108-2	$C_4H_6O_5$	97676
3113-2		
3113-3	$\cdot 2Na$	138090
2990-1	C_4H_7NO	460071
3264-4	$C_4H_7NO_4$	56848
3366		
3142-2	$C_4H_7N_3O$	60275
3387		
3016	$C_4H_7N_4O_4$	3844664
3070-3	$C_4H_8NO_3$	3922767

MOLECULAR FORMULA INDEX

ABSTRACTS 59,749 TO 73,877

Each column in this Molecular Formula Index is consecutively numbered. There are 697 columns in this annual cumulation. For greater clarity and convenience, the molecular formulas appear at the left hand side of the columns, followed by one or more periods, followed by the compound's address which consists of the abstract number and the line number. Furthermore, the C-H headings have been indented. When a C-H grouping is continued in the next column, its heading is repeated at the top of that column (see e.g. top line in column 40).

A hyphen (-) separates a compound number from its abstract number. Commas (,) separate compound numbers appearing in the same abstract, as e.g. in second line, column 40. Diagonal lines (/) separate abstract numbers. The last abstract-compound number listed for any molecular formula is never followed by a (/) or (-) or (,) and the appearance of any punctuation mark at the end of a line indicates that additional abstract-compound numbers for that formula will follow in the next line, column, or page.

The abstract-compound numbers pertaining to those molecular formulas which were corrected by the author after the publication of the abstract to his article, are followed by @ (meaning addition: compound originally not indexed) or * (meaning correction: compound originally indexed under erroneous molecular formula).

Index Chemicus Annual Cumulative July 1966
ISI

INORGANIC

AG AL O2..........71526-1
AG CE5 S8........67348-10
AG CS O..........73238-4
AG CS2 F6 K.......73237-1
AG DY S2..........67836-4
AG DY5 S8.........67348-8
AG ER S2..........67836-7
AG F2 N O4 S2....59896-1
AG GD S2..........67836-2
AG GD5 S8.........67348-6
AG H8 N7 O2 P2 S.72943-4
AG HO.............71993-3
AG HO S2..........67836-6
AG K O............73238-3
AG LA5 S8.........67348-9
AG LI2 SN.........69739-3
AG ND5 S8.........67348-4
AG O18 V7.........65688-1
AG PR5 S8........67348-11
AG S2 SM..........67836-1
AG S2 TB..........67836-3
AG S2 TM..........67836-8
AG S2 Y...........67836-5
AG S2 YB..........67836-9
AG S8 TB5.........67348-7
AG SM5 S8.........67348-5
AG2 GE4 O9........60099-2
AG2 H2 O6 TE2....61281-1
AG2 H4 N4 O6 S3..63281-3
AG2 HO............71993-2
AG2 LI SN.........69739-6
AG2 O3 SE.........69054-1
AG2 O3 TE.........69054-2
AG2 S.............69054-3
AG2 SE............69054-4
AG2 TE............69054-5
AG2 Y.............60682-1
AG3 AU S2.........69054-7
AG3 BR S..........62253-1
AG3 GE7 H O16....60099-3
AG3 I S...........62253-2
AG3 N O3 S........60057-1
AG4 GE9 O20.......60099-1
AG4 H2 N4 O6 S3..63281-4
AG4 N O3 P........64619-1
AG5 O15 P5........63691-2
AG6 O8 V2.........65927-1
AG6 O18 P6........63691-5
AG6 O22 TE V4 W2.70004-2
AG6 O28 V10.......65927-2
AG7 H5 O8 TE.....61281-2
AG7 HO2...........71993-1
AL AS12 N3 S9....60787-11
AL B H6...........65507-2
AL B2 H9..........65507-1
AL B2 MN2.........69065-1
AL BR TE..........71174-2
AL BR6 CL3 O2 P2.67299-6
AL BR9 O2 P2......67299-7
AL C6 CL4 O6 TC.63794-21
AL CD IN O4.......63690-9
AL CL H2..........72532-1
AL CL TE..........71174-1
AL CL4 H..........64142-6
AL CL7 ZR.........67722-1
AL CO IN O4.......63690-4
AL CU IN O4.......63690-6
AL D4 H2 LI3......64053-2
AL F O............69343-1
AL H O2...........71526-2
AL H3 O3..........70323-2
AL H6 LI3.........64053-1
AL HO2............67238-1
AL I TE...........71174-3
AL IN MG O4....63690-1.2
AL IN MN O4.......63690-3
AL IN NI O4.......63690-5
AL IN O4 ZN....63690-7.8
AL NB11 O29.......73232-1
AL RU.............68179-1
AL2 CL10 ZR.......67722-2
AL2 EU O4.........67831-1
AL2 FE H8.........70183-1
AL2 FE2 H2 O26 P4 S U2..71243-1
AL2 GD4 O9........61958-1
AL2 H8 MG.........71271-1
AL2 I6 O21........64210-1
AL2 K4 O9 SI2....62353-1
AL2 O9 SM4........61958-2
AL2 O9 TB4........66233-4
AL2 O9 YB4........66233-6
AL2 RU............68179-3
AL3 BR3 H6 O6...65914-12
AL3 CE............67236-2
AL3 CL3 H6 O6....65914-2
AL3 CL13 ZR.......67722-3
AL3 F H K O11 SI3.71338-3
AL3 F2 K O10 SI3.71338-5
AL3 F18 K6 LI3...72918-1
AL3 H2 K O12 SI3.71338-1
AL3 H12 LI NB....62251-6
AL3 H12 NB........62251-5
AL3 LA............67236-1
AL3 ND............67236-4
AL3 PR............67236-3
AL3 RU2...........68179-2
AL4 B12 BE4 CS O28......66598-1
AL4 CL5 K O4......66301-3
AL4 CL5 LI O4....66301-1
AL4 CL5 NA O4....66301-2
AL4 CL5 O4 RB....66301-4
AL4 CL16 ZR.......67722-4
AL5 BR3 H12 O12.65914-14
AL5 CL3 H12 O12..65914-4
AL5 H20 LI NB2...62251-2
AL5 H20 NB2.......62251-1
AL5 RU2...........68179-4
AL6 CL3 H15 O15..65914-5
AL6 F H3 K2 O23 SI6......71338-2
AL6 F3 H K2 O21 SI6......71338-4
AL7 H28 LI NB2...62251-4
AL7 H28 NB2.......62251-3
AL8 F48 K9 LI15..72918-3
AL9 BR6 H21 O21.65914-13
AL9 CL6 H21 O21..65914-3
AL10 O24 TB6......66233-5
AL10 O24 Y6.......66233-3
AL10 O24 YB6......66233-7
AL11 BR3 O30....65914-15
AL11 CL3 H30 O30.65914-7
AL11 F66 K15 LI18.72918-2
AL12 EU O19.......67831-2
AL13 BR15 H24 O24.65914-11
AL13 CL15 H24 O24.65914-1
AL13 RU4..........63672-1
AL15 HF6 NI8......65410-8
AL15 NI8 ZR6......65410-7
AL16 HF6 NI7......70344-6
AL17 BR3 H16 O32..65914-17
AL17 CL3 H16 O32.65914-9
AL22 LA2 O36......66233-1
AL22 ND2 O36......64315-1
AL22 O36 SM2......66233-2
AL24 BR6 H44 O55..65914-16
AL24 CL6 H44 O55.65914-8
AL29 CL9 H78 O78.65914-6
AL136 BR3 H135 O270.....65914-18
AL136 CL3 H135 O270.....65914-10
AM F2 O2 RB.......59908-3
AM4 O18 XE3.......73194-1
AS B..............68454-1
AS CD2 CL2........69049-1
AS CD3 CL3........69049-3
AS D..............70892-3
AS D2.............70892-4
AS F3 O...........65711-1
AS F5.............73136-5
AS F5 O2 S........63853-8
AS F8 N O.........68148-2
AS H..............70892-1
AS H O8 P2........71532-1
AS H2.............70892-2
AS S3 SB..........64654-1
AS2 B2 BA2 O10...71095-2
AS2 B13...........68454-2
AS2 BA4 O9........61923-7
AS2 BR3 HG4.......62476-1
AS2 CA4 O9........61923-2
AS2 CD4 CL3.......69049-2
AS2 FE4 O11.......62252-1
AS2 HG4 I3........62476-2
AS2 I4............69048-1
AS2 MO............60310-3
AS2 O7 TH.........71157-2
AS2 O9 SR4........61923-5
AS2 O10 PB2 U....60803-1
AS2 TH............63449-1
AS2 V3............73416-2
AS2 W.............60310-1
AS3 MO2..60310-4/60325-1
AS3 V5............73416-1
AS3 W2...60310-2/60325-2
AS4 H N S3........60787-1
AS4 H12 N4 S4....60787-8
AS4 H16 N6 S3....60787-5
AS4 LI N S3......60787-12
AS4 MO5...........60310-5
AS4 O12 TH........71157-1
AS4 O12 TL4.......60058-1
AS4 O16 TH3.......71157-4
AS6 O23 TH4.......71157-3
AS7 RE3...........63450-1
AS7 TC3...........63450-2
AS8 BA N2 S6......60787-9
AS8 H42 MG N16 S6.60787-7
AS8 H48 BA N18 S6.60787-6
AS8 MG N2 S6....60787-10
AT BR.............68075-4
AT CL.............68075-5
AT H..............68075-1
AT I..............68075-3
AU CU TE4.........65478-1
AU LI2 SN.........69739-4
AU NA2 O6 S4......69054-6
AU2 Y.............60682-2
B BR F6...........70164-1
B BR3 I3 P.....65460-2.4
B BR6 P...........65460-5
B CL3 I3 P........65460-3
B CO3 O5..........72935-1
B CR..............71086-3
B CR2.............71086-1
B CS H4...........70842-2
B F H2 O2.........62254-1
B F2 H3 O2........62254-2
B F3..............73136-3
B F3 H2 K N.......72535-2
B F3 H2 N NA......72535-1
B F3 H4 O2........62254-3
B F3 H5 P SI......66471-1
B F4 H5 O2........62254-4
B F4 N O..........61838-2
B H3 O2...........73549-1
B H3 O6 P4........68116-1
B H4 K O3.........73549-2
B H4 RB...........70842-1
B H8 N2 NA........60226-3
B H12 N4 NA.......60226-1
B H28 N12 NA......60226-4
B I6 P............65460-1
B K S3............61009-3
B NA S3...........61009-1
B2 BA2 O10 P2....71095-1
B2 BR F9..........73630-2
B2 BR2 S3.........70187-1
B2 CL2 S3.........70187-2
B2 CR.............71086-5
B2 CR3...........*71086-2

3

B2 CU F8 H18 N6..68091-4
B2 CU2 O5.........59802-1
B2 F8 H15 N7 RU..62582-3
B2 H2 O3..........70779-1
B2 H6 O6 P4......67550-1/68116-2
B2 H8 N2..........73569-1
B2 H9 P...........61497-1
B2 H10 N4 O4......64656-1
B2 H12 N2 NA2....60226-2
B2 H18 N6 NI O4..66643-3
B2 H26 N6 NI......60744-1
B2 I2 S3.68663-3/70187-3
B2 K2 S5..........61009-4
B2 NA2 S5.........61009-2
B2 PB2 S5.........65681-1
B2 W..............62379-1
B3 H4 N O5........60605-1
B3 H8 N O7........60605-2
B3 H9 O6 P4.......68116-3
B3 H12 N3.........73569-2
B4 BR EU H4.......70351-3
B4 CR3............71086-4
B4 CU3 O9.........59802-2
B4 D2 H8..........65555-1
B4 H4 O4..........65770-1
B4 H9 N O9........60605-3
B4 H16 N4.........73569-3
B4 NA2 O7.........70807-3
B5 CL H8..........73587-1
B5 H8 N O10.......60605-4
B5 H11............65555-2
B5 H20 N5.........73569-4
B6 BA.............71553-3
B6 CA.............71553-1
B6 CO12 RE11......65401-9
B6 CO21 SC2.......65401-1
B6 FE2 RE21.......65401-4
B6 FE6 RE17.......65401-5
B6 FE11 RE12......65401-6
B6 FE13 RE8.......65401-8
B6 FE13 RE10......65401-7
B6 H4 O7..........65770-2
B6 H30 MG2 O26...61201-1
B6 MN11 RE12......65401-3
B6 NA2 O9.........70807-2
B6 NI11 RE12....65401-10
B6 NI21 SC2.......65401-2
B6 SR.............71553-2
B8 NA2 O13........70807-1
B9 H16 N..........64415-3
B10 CL8 CS H2 N3.64871-9
B10 CL8 CS2 H2 S2.64871-13
B10 CL8 CS2 N5..64871-11
B10 CL8 CS2 N6..64871-12
B10 CL8 H3 N3....64871-8
B10 CL8 N4........64871-7
B10 CS H16 N......64415-1
B10 CS2 H14 O....64415-2
B10 H8 N4.........64871-1
B10 H12 N2 TL2..64871-20
B10 H14...........70634-1
B10 H14 N2........64871-5
B10 H14 N3 TL...64871-21
B10 H15 N3.......64871-18
B10 I8 N4........64871-14
B12 BR4 CS2 H8...67533-1
B12 BR5 CS2 H7...67533-2
B12 BR10 CS2 H2 O2.......67533-13
B12 BR10 CS2 N6.67533-11
B12 BR11 CS2 N3.67533-10
B12 NI............71510-1
B20 H18 K4 U......61248-3
B20 H20...........71864-1
B20 H22 O.........71864-2
BA BE F4..........63995-2
BA BR2 F4 O4......62012-6
BA CL F...........59805-2
BA CL2 F4 O4......62012-5
BA CL4 CS2........73774-1
BA CR2 H8 N2 O8.71096-10
BA CR2 K2 O8......71096-9
BA CR2 O8 RB2...71096-11
BA CR2 O8 TL2...71096-12
BA DY2 O4.........72181-4
BA ER2 O4.........68367-3
BA EU2 O4.........72181-2
BA GD O4..........72181-3
BA H4 N4 O6 S3...63281-5
BA HO2 O4.68367-2/72181-5
BA LU2 O4.68367-6/72181-7
BA ND2 O4.........72181-1
BA O4 TM2.........68367-4
BA O4 W...........70217-1
BA O4 W3..........70217-4
BA O4 Y2..........68367-1
BA O4 YB2.68367-5/72181-6
BA O7 W2..........70217-2
BA O13 W4.........70217-5
BA SC2 O4.........62360-1
BA2 CL6 CS2.......73774-2
BA2 CU O6 W.......66453-1
BA3 CD N6 O12....68098-3
BA3 CU NB2 O9....66453-2
BA3 CU O9 TA2....66453-3
BA3 DY2 O9 W....66453-16
BA3 GE3 O9.......64248-19
BA3 HO2 O9 W....66453-15
BA3 IN2 MO O9...66453-12
BA3 IN2 O9 W....66453-13
BA3 IR O9 RU2....65442-1
BA3 LI MO NB O9..66453-9
BA3 LI MO O9 TA.66453-10
BA3 LI NB O9 W...66453-7
BA3 LI O9 TA W...66453-8
BA3 MG O9 RU2....65442-4
BA3 MN O9 RU2....65442-2
BA3 N2............70041-2
BA3 N4............70041-1
BA3 NI O9 RU2....65442-3
BA3 O8 P2.........63137-1
BA3 O9 SC2 W....66453-14
BA3 O9 SC4 62360-2

4

BA3 O18 P6........63691-6
BA3 O22 TE V4 W2.70004-3
BA3 O24 P6 TE2...61435-2
BA4 LI NB3 O12..66453-17
BA4 NA NB3 O12..66453-18
BA4 O9 P2.........61923-6
BA4 O16 SI6.......73148-3
BA4 O36 SI12 SN3 TI......60805-1
BA5 NA2 O15 W3..66453-11
BA5 O18 RU6 SR...65441-1
BA5 O30 P10.......63691-3
BA5 O41 W12.......70217-3
BA5 SI3...........66607-1
BA8 MO18 O82 P2 TE5.....61435-6
BA9 N10 OS3.......70776-3
BA9 N10 RE3.......70776-1
BE CL4 K2.........62344-1
BE CL4 NA2........62344-2
BE CS2 O8 S2......61229-1
BE CU F4..........63995-3
BE F4 H2..........63995-1
BE H4 N2..........64738-2
BE3 N2............64738-4
BE15 CO8 HF6......65410-3
BE15 CO8 ZN6......65410-1
BE15 CU8 HF6......70344-3
BE15 HF6 NI8......65410-4
BE15 HF6 PD8......70344-2
BE15 NB6 NI8......65410-5
BE15 NI8 TA6......65410-6
BE15 NI8 ZR6......65410-2
BE15 PD8 ZR6......70344-1
BI CE.............60180-1
BI DY.............60180-7
BI ER.............60180-8
BI GD.............60180-5
BI ND.............60180-3
BI O4 U...........65186-3
BI PR.............60180-2
BI SM.............60180-4
BI TB.............60180-6
BI2 CL7 RB........66285-2
BI2 IR............63652-6
BI2 O3............60742-6
BI2 O6 U..........65186-1
BI2 O6 W..........72545-1
BI2 O12 S3........60742-1
BI2 RH............63652-5
BI3 IR............63652-4
BI3 NI............63652-1
BI3 RH............63652-3
BI3 RU............63652-2
BI4 O9 U..........65186-2
BI6 O18 S3........60742-3
BI6 O27 S6........60742-2
BI8 O21 S3........60742-4
BI10 O24 S3.......60742-5
BR CE S...........69123-3
BR CO H15 N7 O6..73301-2
BR CR H10 O9 S...61307-2
BR CR O...........71119-1
BR CR S...........69741-1
BR CS O4 SE2......72903-9
BR F2 K O2........62012-4
BR F2 NA O2.......62012-3
BR F2 P S.........63793-2
BR F3 TI..........63566-1
BR FE N3..........69066-4
BR GE2 H5.........68743-2
BR H N RE.........60740-5
BR H6 K SI2.......68742-1
BR K O4 SE2.......72903-8
BR K4 N5 O10 PB..73768-6
BR K5 N8 O16 PB2.73768-5
BR LA S...........69123-2
BR LI O2 SE.......72903-2
BR NA O2 SE.......72903-3
BR O2 PA..........65069-3
BR O3 RE..........65769-2
BR SE TL..........66447-1
BR2 CD H8 N4......61811-1
BR2 CL6 MG O2 P2.61047-1
BR2 CL9 MG O3 P3.61047-2
BR2 CO H15 N6 O..61463-1
BR2 CU H18 N6....68091-2
BR2 EU............70351-1
BR2 F10 GE........73630-1
BR2 H IN O7.......73315-2
BR2 H5 N2 RE......60740-4
BR2 H6 N2 ZN......60741-5
BR2 H8 N4 ZN......61811-2
BR2 H15 N7 RU....62582-1
BR2 H18 MN N6....73150-1
BR2 HG2 SB........69047-1
BR2 O TH..........69155-7
BR3 CL5 O P SB...67299-5
BR3 CO H15 N5...62973-13
BR3 CS PB.........72186-1
BR3 GA O9.........73315-1
BR3 H12 N4 RE....60740-2
BR3 O PA..........65069-2
BR3 O7 P S2.......67299-1
BR3 SB............72283-1
BR4 CL8 MO6.......73738-7
BR4 CL8 W6.......73738-10
BR4 CS O RE......61300-4/63490-1
BR4 H MO..........71919-5
BR4 K SM.........65863-15
BR4 LI O P........67299-8
BR4 NA SM........65863-14
BR4 P.............68506-1
BR4 RB SM........65863-16
BR4 TA............59902-1
BR5 CO CS3........69276-1
BR5 CO2 H30 N9 O2.62032-1
BR5 H2 MO.........71919-6
BR5 H2 MO O.......63811-1
BR5 H2 O W........63811-2
BR5 H8 N3 O RU...68768-4
BR5 K2 N O RU...68768-5/70941-2
BR5 K2 SM.........65863-5

5

BR5 NA2 SM........65863-4
BR5 PA............65069-1
BR5 RB2 SM........65863-6
BR6 CL4 O2 P2 SN.67299-3
BR6 CL4 O2 P2 TI.67299-4
BR6 CL8 H2 MO6...73738-2
BR6 CL8 H2 W6....73738-5
BR6 CS2 RE........61300-5
BR6 CS4 PB........72186-2
BR6 K3 SM........65863-10
BR8 CL RE3........65769-1
BR8 CL8 W6.......67240-12
BR8 NB3...........71011-1
BR10 K4 O RU2....68284-2
BR10 N5 P5........60886-1
BR10 O2 P2 SN....67299-2
BR11 GA9 SB.......66386-2
BR12 MO6..........67240-2
BR12 W6...........67240-3
BR14 NB TA5.......71013-1
BR14 NB2 TA4......71013-2
BR14 NB6..........67240-6
BR14 TA6..59901-1/59902-3/67240-8
BR15 TA6.........67240-11
BR15 W3...........68061-2
CA CS N3 O6.......71087-4
CA CS5 N7 O14....71087-5
CA DY2 O4.........62394-1
CA HF4 O9.........62768-1
CA HG N4 O2 P2 S2.69068-2
CA K2 N4 O8.......71087-2
CA K4 O18 P6....72877-10
CA LI N...........70545-1
CA N2 SI..........65132-1
CA N3 O6 RB......71087-3
CA N4 NA2 O8.....71087-1
CA NA4 O18 P6....72877-6
CA O6 P2..........72877-1
CA O8 P2 SN.......71532-9
CA O8 P2 TI......71532-10
CA O8 P2 U........61198-1
CA2 NA2 O7 SI2...68372-1
CA2 NB O6 V.......68194-1
CA2 O5 YB2........62394-2
CA2 O6 TA V.......68194-2
CA3 GA2 O12 SI3..70844-1
CA3 N2............73137-1
CA3 O6 YB2........62394-3
CA3 O12 SC2 SI3..70844-2
CA3 O16 PB SI4 ZN4......60806-1
CA4 N4 SI.........65132-3
CA4 O9 P2.........61923-1
CA4 O9 V2.........61923-3
CA5 N6 SI2........65132-2
CA5 O12 S SI2....66654-1
CA8 H6 NA2 O22 SI5......65479-1
CA8 O18 SI5.......65910-1
CD CS2 I4.72184-2/72185-3
CD GA IN O4......63690-19
CD GE LI2.........69740-2
CD GE O3.........64248-18
CD H8 I2 N4.......61811-3
CD HG O5 SE.......63857-4
CD HG2 O6 SE......63857-6
CD I3 K...........72185-1
CD I4 K2..........72185-2
CD I4 NA2.........72184-1
CD IN2 O4........63690-23
CD K4 O18 P6....72877-12
CD LI2 SN.........69740-5
CD NA4 O18 P6....72877-8
CD O3 TE..........70843-4
CD O6 P2..........72877-3
CD TE.............70843-2
CD2 F7 K3.........73868-1
CD2 GE7 O16......64248-12
CD2 HG3 O11 SE2..63857-5
CD4 I3 SB2........63693-1
CE CL S...........69123-5
CE CS3 F6.........66990-3
CE CS4 O28 W8....70357-5
CE F4 K...........66990-2
CE F5 H4 N........66303-3
CE F6 H8 N2.......66303-2
CE F6 K3..........66990-1
CE F7 H12 N3......66303-1
CE GE O4.........64248-14
CE H16 N4 O28 W8.70357-2
CE K O12 TA4......65461-1
CE K4 O28 W8......70357-3
CE NA4 O28 W8....70357-1
CE NB3 O9.........73725-4
CE O3 TI..........69043-2
CE O8 TE2........72368-15
CE O28 RB4 W8....70357-4
CE O30 W10........68076-1
CE PO.............71468-3
CE SI............73302-12
CE SI2............73302-7
CE2 EU S4........65140-10
CE2 K6 NB4 O16...65424-6
CE2 O9 TE3.......72368-14
CE2 S4 SM.........65140-2
CE2 S4 YB........65140-18
CE2 SI............65387-1
CE3 NB O7.........73725-3
CE3 NI6 SI2.......69061-1
CE3 SI2..........73302-11
CE4 GE3 SI2.......66387-2
CE5 SI3..........73302-10
CE6 SI5...........65387-2
CL CO H12 N6 O4..66481-1
CL CO H15 N5 O4 S.62973-10
CL CO H15 N7 O6..73301-1
CL CR H10 O9 S...61307-3

6

CL CR N2 O2.......60785-1
CL CS F2..........61837-1
CL CS O4..........66591-6
CL CS O4 SE2......72903-7
CL F2 K...........61837-3
CL F2 K O2........62012-2
CL F2 N O S.......73196-2
CL F2 N O4 S2....68074-1
CL F2 N S.........73196-4
CL F2 NA O2.......62012-1
CL F2 P S.........63793-1
CL F2 RB..........61837-2
CL F15 P5 RE......71645-1
CL FE N3..........69066-3
CL GE2 H5.........68743-1
CL H3 O6 S ZR....72170-2
CL H6 I N2 PT....66996-1
CL H6 MO N3 O....67052-2
CL H6 N5 O5 PT..62361-11
CL H7 MO N3 O....67052-1
CL H8 N3 O4 PT...66996-2
CL H9 N4 O3 PT..66996-3/70205-1
CL H12 AG N3 O3..65937-1
CL H12 N4 O2 RE.72640-21
CL HG2 N O4.......60677-3
CL I S............71023-3
CL IR O2..........65768-1
CL K O4 SE2.......72903-6
CL K4 N5 O10 PB..73768-4
CL LA S...........69123-4
CL LI O2 SE.......72903-1
CL MN O3..........65237-1
CL N O3 S.........71023-1
CL N3 S4..........59910-1
CL O2 PA..........69692-1
CL O4 TL..........66591-3
CL2 CO F H15 N5..62973-9
CL2 CO H8 N4......69703-2
CL2 CO H15 I N5.62973-14
CL2 CO H15 N6 O2..62973-1.2/70019-2
CL2 CO H15 N6 O3.62973-7
CL2 CO H16 N5 O4 S.......62973-11
CL2 CO H24 N12...69703-1
CL2 CR H8 N4......69060-1
CL2 CR H12 O15 P.70781-1
CL2 CU H18 N6....68091-1
CL2 CU H18 N6 O8.68091-5
CL2 F2 TI.........62757-1
CL2 F2 ZR.........73316-1
CL2 GE............66232-5
CL2 H3 N O11......61991-1
CL2 H6 N2 ZN......60741-4
CL2 H6 N4 O4 PT..70210-1
CL2 H10 O14 V....61855-1
CL2 H12 N4 O2 PT..72541-10.13
CL2 H15 I IR N5..73186-5
CL2 H15 IR N6 O3.73186-6
CL2 H16 IR N5 O9.73186-4
CL2 H18 N6 O4 S TI......62257-9
CL2 HG3 S2........68480-1
CL2 K2 N4 O8 PT.62361-10
CL2 K6 N6 O12 PB.73768-3
CL2 MO O2.........63579-5
CL2 N P...........73188-1
CL2 N2 S3.........73198-1
CL2 NB O..........67735-6
CL2 O TH..........69155-8
CL2 O U...........66592-5
CL2 XE............65192-1
CL3 CO H15 N5..62973-12/70019-1
CL3 CO H17 N5 O..62973-3
CL3 CS CU.........64410-1
CL3 CS2 F3 PT....72642-3
CL3 D O U.........66592-2
CL3 F3 K2 PT......72642-1
CL3 F3 PT RB2....72642-2
CL3 F3 W..........70798-2
CL3 FE RB.........63852-1
CL3 H O U.........66592-1
CL3 H2 N U.......68744-18
CL3 H3 N2 SN......65349-1
CL3 H4 N O3 RU...61230-5
CL3 H6 N3 O2 PT..70210-7
CL3 H7 N4 SN......65349-2
CL3 H11 N6 SN....65349-3
CL3 H12 N5 O RU..61230-3
CL3 IR............65768-2
CL3 K O SE........68740-6
CL3 MO N..........60785-3
CL3 MO S2.........60733-1
CL3 N P...........67051-7
CL3 N W...........60785-4
CL3 O PA..........69692-3
CL3 O TA..........65427-5
CL3 O W...........68488-1
CL3 PT SN3........63485-4
CL3 TI............64384-3
CL4 CO2 H30 N12 O2......71019-1
CL4 CS HO.........71919-3
CL4 CS O TA.......65427-3
CL4 CS2 NB O......67735-4
CL4 CS2 NP O2....68288-4
CL4 CS3 NP O2....68288-3
CL4 CU5 H6 O6....66981-1
CL4 F8 NI P4......68643-1
CL4 FE RB2........63852-2
CL4 H MO..........71919-2
CL4 H4 N O TA....65427-4
CL4 H5 N3 O2 PT..70210-2
CL4 H6 N2 PT......63882-1
CL4 H6 N2 U.......68744-7
CL4 H8 N4 SN......65349-4
CL4 H8 O16 S2 ZR4.70854-4
CL4 H9 N3 PT......63882-2
CL4 H12 N4 PT....63882-3
CL4 H12 N4 U......68744-6
CL4 H12 N6 SN....65349-5
CL4 H16 N8 SN....65349-6
CL4 H24 N8 U......68744-5
CL4 H52 O6........62584-2

MACHINE LANG.

4. Machine Language Representation

A. Punched Tabulating Cards

   1. The Storage Medium

   2. The Coding

B. Paper Tape

C. Magnetic Media

D. Character Readers

   1. Handwritten Characters

   2. Page Readers

E. Machine Representative of Natural Text

F. ASCII Code

REFERENCES

4

Machine-language representation

For machine retrieval or processing, some numeric, alphabetic, or special symbols must eventually be transformed to some physical representation—such as holes in a card or magnetized spots on a magnetic tape—in order that they can be stored and interpreted by the equipment. Most of the equipment in use today has a limited capability for representing and working with symbolic or textual information. Most of the methods for representing information on the various storage media (punched cards, punched tape, magnetic tape, etc.) were developed and adapted by the manufacturers of equipment that utilized these media. For example, both IBM and Remington Rand had developed their own separate card and code designs to go with their equipment; the cards are not compatible, and each of the coding systems uses only a small portion of the coding potential of the cards. There have been some attempts to develop a common or standard coding system for mechanized information processing equipment, but no definite standards have yet been established. The following sections describe the storage medium and coding systems used for punched tabulating card (tab card), paper tape, and magnetic tape systems.

A. PUNCHED TABULATING CARDS

1. The Storage Medium

IBM or Remington Rand punched tabulating cards are used for virtually all of the mechanized card-handling systems in the United States, with IBM supplying most of the installations.[1] Each of these card types (see Figs. 4-1 and 4-2) comes in a variety of sizes (normally 51-, 60-, 66-, and 80-column cards for IBM; and 32-, 40-, 54-, 66-, and 90-column cards for Remington Rand), but most organizations use the largest size possible. The smaller sizes are useful for some special applications, as will be mentioned later. The cards are available in a variety of colors, and are usually preprinted for the user's particular requirement. The characters (numbers, alphabetic symbols, and some special symbols), which are used as the restricted languages of the various Remington Rand and IBM punched card machines, are shown in Figs. 4-1 and 4-2.

The IBM card is most commonly used in such a way that each of the vertical columns on a card represents a single character. Consequently, the largest card size (80 columns) will permit 80 characters of information to be stored on a card. The Remington Rand card normally allows a maximum of 90 characters to be stored on a single card. The Remington Rand equipment normally operates with a smaller alphabet of characters than the IBM equipment. That is, it recognizes a smaller number of different symbols. Generally, the card equipment will accept and interpret the cards only in the manner of coding shown in Figs. 4-1 and 4-2. Arbitrary coding schemes devised by the card users

[1] Throughout this book, the IBM and Remington Rand type of card will often be referred to as a tab (tabulating) card, an expression in common usage.

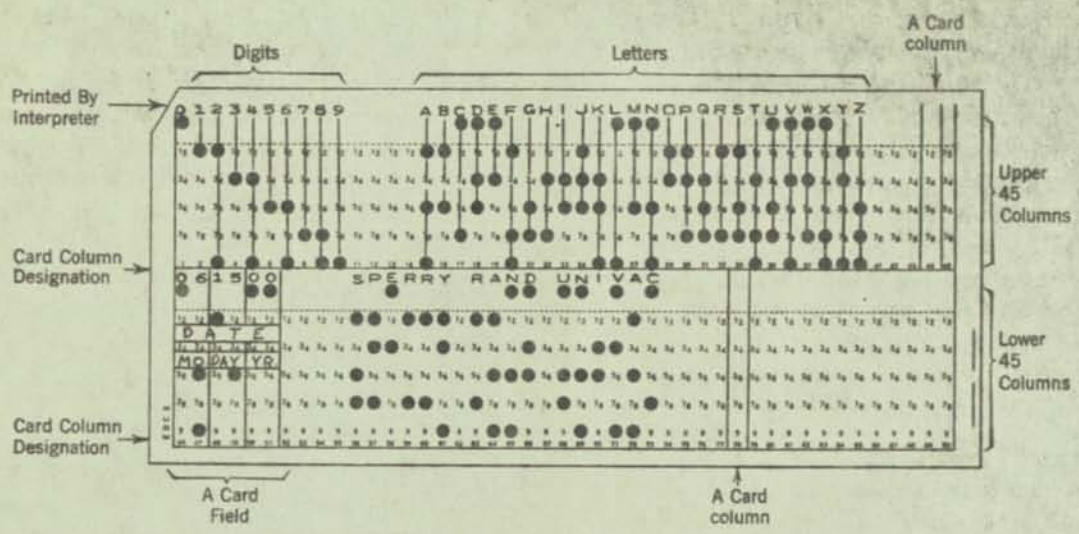


FIG. 4-1 Remington Rand 90-column card.

will usually be meaningless to the card-handling equipment.

From a practical standpoint the cards can tolerate a certain degree of rough handling, but do not work well when they have been folded, creased, warped, or stapled. Stapled or mutilated cards may jam the machines, or result in a misinterpretation of the data punched in them. Special card-editing equipment is available for the primary purpose of reading cards which have been handled by the public, in order to extract the illegible or manhandled cards. Another problem is that the cards gradually become frayed or damaged during repeated machine handling and may occasionally have to be regenerated. Some card-handling problems may be aggravated by temperature or humidity effects; consequently, the cards should be stored in a cool, dry atmosphere.

Some card stock is available which is perforated or scored, to permit a portion of the card to be detached for subsequent card-handling operations. This detached portion might be used, for example, as a return stub for billing purposes, with a per-

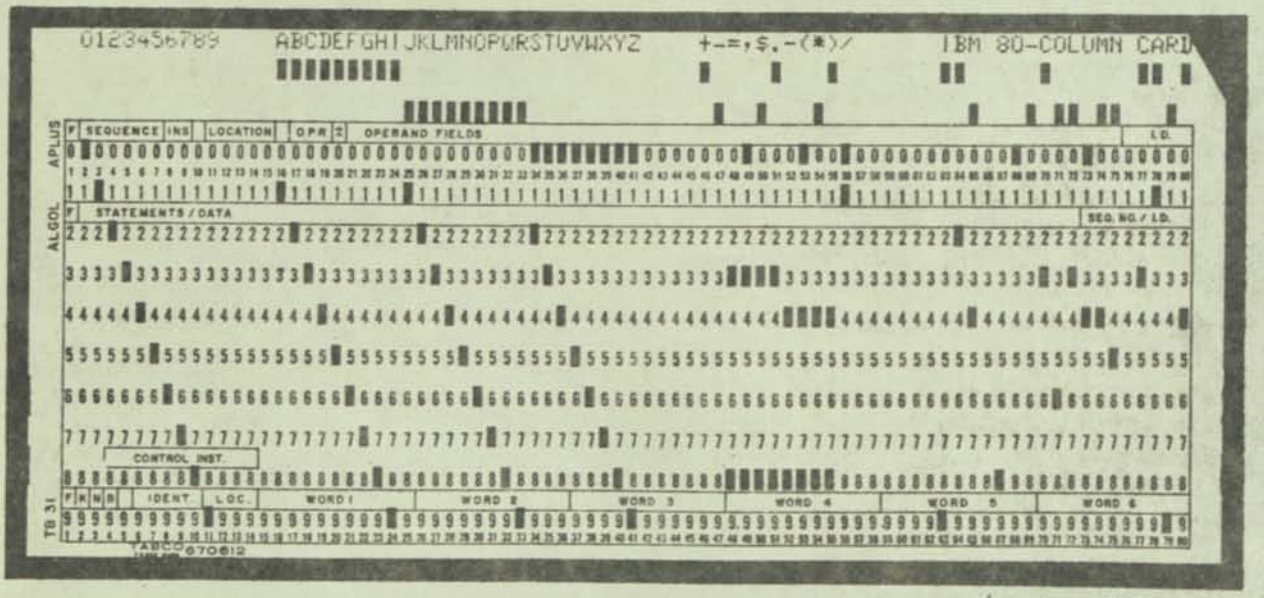


FIG. 4-2 IBM 80-column card.

forated 80-column card sent out as a bill and the 21-column stub returned by the customer with the payment. The stub (previously punched with the customer's account number and amount due) could serve as the input card for the machine to post a credit to the customer's account. Equipment modifications are available to permit the handling of any of the card sizes, but most operating systems reproduce the information from the smaller card into the largest card size as soon as the smaller card is received.[2]

Symbols are usually entered on the card by one or more of these four methods:

1. Manual keypunching
2. Automatic punching
3. Mark-sense cards
4. Pre-scored cards

In most situations, the cards are manually prepared by an operator who reads a source document and punches the data into a card, using a key punch machine (see Fig. 6-1). That is, the information is transcribed from a source document into a card. Less expensive card punches with reduced capability are available. In some cases, cards are automatically punched by computer output devices or tabulating equipment as an output, or temporary storage of the result of some machine operations.

Mark-sense cards and their associated equipment provide a capability for interpreting information that has been marked on the cards with a special pencil. Equipment is available to read these penciled marks and punch the data into the same card or into new cards. Mark-sense cards can be used in some cases to reduce the amount of manual keypunching required, but the system has found widest use where some data must be entered at the point of origin (e.g., recording physical inventories, survey results, and utility meter readings). Each marking position on the card is 3 columns wide, thus restricting the card to a maximum of 27 mark-sense characters, which may be a disadvantage in some applications. In general, any number of marking positions from 1 to 27 can be placed on a card, and in any desired position on the card. Many people with extensive data processing experience tend to avoid use of the mark-sensing technique except in those instances where there are several outstanding advantages. In general, users have indicated satisfaction with the results only under the following circumstances: there were a small number of digits to be entered; there was no pressure or requirement for speed at the time of entry; there were good physical working conditions; and the entries were to be made by clerical personnel. Most decisions to use mark-sense cards are based upon possible savings in time or cost. A cost advantage may be realized because manual keypunching and verifying is eliminated, and a time saving may be realized because the cards can be processed immediately upon their receipt, without waiting for keypunching and verifying. There does not appear to be any significant difference in the accuracies achieved by recording with mark sensing rather than by data recording and subsequent keypunching.

Pre-scored cards such as Port-a-Punch (see Fig. 4-3) permit data to be punched into the card by pushing out the perforated holes with a pencil or other probe. The pre-scored holes are positioned with a 1-column spacing between each pre-scored column. Consequently, only 40 characters, each separated by a 1-column width, can be punched into a card. The usual practice is to reproduce the information from the pre-scored cards into a standard card as soon as the card is received. This con-

FIG. 4-3 IBM Port-A-Punch.

[2] Some very small punched tags are available from the Dennison Company for use as retail sales tags, and equipment is available to read these tags and punch the data into standard punched cards.

denses the information on the card, and removes the possibility that additional holes may pop out during subsequent machine handling. These pre-scored cards may also be useful for small, peek-a-boo file systems (see Chapt. 5).

Card sets are also available that permit duplicate papers (e.g., receipts) and punched cards to be prepared at the same time, usually at the information source. A large number of retail charge systems, such as gasoline credit cards, use this type of card set.

2. The Coding

Standardization of the Hole Patterns. As mentioned earlier, the pattern of holes in the cards is meaningless unless some definitions or standards are established by the users. There are 960 hole positions on the IBM card and 540 on the Remington Rand card. With this number of hole positions, a tremendous number of unique hole patterns are possible. However, the equipment manufacturers have adopted their own coding definitions, which utilize only a small percentage of the code possibilities. Most of the equipment recognizes only a small number of hole patterns to be valid, and usually treats any other hole patterns as errors. In some applications, such as the sorting of cards with superimposed coding on the Statistical Sorter (described later), unconventional hole patterns can be used. The restricted number of valid codes and characters (the machine language) designed into the early equipment were generally adequate for the early applications of this equipment. However, the steadily increasing application of tab card and computer equipment to a great variety of information processing tasks has created a need for more flexible coding and a greater number of valid code combinations which can be machine-manipulated. For example, sophisticated automatic programing systems for computers can be used most efficiently when the tab card codes directly represent the symbols of the programing language (e.g., $\geqq$, $\sqrt{\ }$, $\neq$, { }, [ ]). Some code modifications have been proposed for this purpose.[3]

[3] Voorhees, E. A., "Some Thoughts on Reconciling Various Character Set Proposals," *Communications of the Association for Computing Machinery*, Vol. 3, No. 7, pp. 408–409 (July 1960).

Bemer, R. W., "A Proposal for a Generalized Card Code for 256 Characters," *Communications of the Association for Computing Machinery*, Vol. 2, No. 9, pp. 19–23 (September 1959).

In using a set of valid characters established for a card code, there are several methods most often employed for representing information on the card:

Direct coding
Abbreviated coding
Category coding
Category coding with unconventional punching
Superimposed coding

Direct Coding. This is the most common method and the technique that most of the equipment has been designed to utilize. With this method, there is a one-for-one correspondence between the original source data and the characters in the card columns. For example, a card used in a system for automobile registrations may contain the complete spelling of the owner's name and address as well as the vehicle information. Punching the information in this manner will allow the cards to be sorted (by name, address, or license number), or duplicated for use in separate reference files (e.g., by name, address, or license number), or listed by a printer to obtain catalogs, directories, or mailings. Telephone directories, parts catalogs, and name lists are often generated and maintained with cards that use this method of coding.

Abbreviated Coding. In order to make efficient use of the available card columns it may be necessary to abbreviate or condense the spelling of a name, part description, subject term, or other information to be punched. Any of the methods of assigned and derived codes described in Chapt. 3 could be used to obtain these abbreviations. Cards, for example, which contain the information for addressing magazine labels or advertising material often make use of abbreviation techniques.

Category Coding. This technique is commonly employed with tab cards and edge-punched cards to make efficient use of the available coding positions. Category coding requires the use of a reference table to describe which items belong in each category so that the original terms can be correctly assigned to the proper category. The following examples illustrate how various topics might be represented as numbered categories.

Subject Field	*Code*
Physics	01
Chemistry	02
.	.
.	.
.	.
Zoology	99

Geographic Area	*Code*
San Francisco	01
Los Angeles	02
.	.
.	.
.	.
San Diego	99

~~*Uniterm*~~ *Subject Heading*	*Code*
Aerodynamic Heating	1334
Diffraction	2123
Methyl Groups	3024

Other methods of numeric coding were described earlier, in Chapt. 3. A description of various ways to indicate relationships between descriptors coded in different areas of a card has been given by Dekker.[4]

Category Coding with Unconventional Punching. Most punched card equipment recognizes only a small group of punch patterns as being valid codes, and rejects all other patterns as illegal. The processing equipment either interprets these invalid patterns as errors or manipulates the information to provide an incorrect result. Printing or tabulating equipment does not have any characters to correspond to the invalid punches, and key punch machines cannot conveniently be used to punch the invalid patterns into the card. However, some of the sorting equipment can operate with the invalid punches.

Because the use of non-standard punch patterns (e.g., a 7 and an 8 punched in the same column) will provide a greater number of coding possibilities for each card column, the use of the invalid patterns provides a more efficient coding scheme than could be obtained with direct or category coding with conventional punching. In a strict sense, it is possible to uniquely represent 2^{12} or 4096 different topics in one IBM card column by using this technique. However, for convenience the user may wish to employ only those code combinations which utilize 2 or 3 punches per column. This will decrease the number of total coding possibilities, but this number may still be much larger than that obtained with conventional coding. Unconventional punching might be used, for example, to permit a single column to represent many different animals, such as:

Animal	Rows Punched in a Column
Horses	3 and 4
Cows	3 and 5
Sheep	1 and 6
.	.
.	.
.	.
Cats	7 and 8

The codes might also be assigned in a systematic manner to represent a hierarchical structure in a single column. For example:

Subject	Rows Punched in a Column
Airplanes, all types	1
Airplanes, jet	1 and 2
Airplanes, turboprop	1 and 3
Airplanes, propeller	1 and 4
Airplanes, helicopters	1 and 5
Ground vehicles, all types	2
Ground vehicles, 4 axles	2 and 3
Ground vehicles, 6 axles	2 and 4
Ground vehicles, 8 axles	2 and 5
Ground vehicles, more than 8 axles	2 and 6
Ground vehicles, tracked	2 and 7
Boats, all types	3
etc.	etc.

Superimposed Coding. It is possible to superimpose descriptions or indexing information into fixed fields of a tab card.[5] This process will usually generate some invalid punch combinations and introduce the inconveniences of invalid punch combinations men-

[4] Dekker, J., "Interfiling of Descriptors for Row-by-Row Coded Punched Card Machine Literature Searching Systems," Chapt. 8 in *Information Retrieval and Machine Translation*, Part I, A. Kent, editor (Interscience Publishers, New York, 1960).

[5] Anon., "Superimposable Punched Cards as a Means of Reference to Periodicals," ~~*Unesco Bulletin for Libraries*~~, Vol. 12, No. 10, pp. 228–229 (October 1958).

Weil, B. H., and E. A. Clapp, "The Adaptation of the ASM-SLA Metallurgical Literature Classification for Use in a Machine-Sorted Punched Card System," Chapt. 20 in *Information Systems in Documentation*, J. H. Shera et al., editors (Interscience Publishers, New York, 1957).

tioned earlier. The use of the superimposed punching will normally restrict the card-handling operations to the use of the IBM Statistical Sorter (described later, in Chapt. 6).

6. PAPER TAPE

Punched paper tape is available in many different sizes and is used with many different code patterns. The code may be represented by holes that are completely punched out (chad tape) or partially punched out (chadless tape) or printed but not punched (e.g., Burroughs Electrographic tape). The holes may be sensed with mechanical feelers (for chad and chadless tape), sensed photoelectrically by light beams transmitted through the holes (chad tape), or sensed by beams reflected from the printed spot (Electrographic tape). Tape that is sensed by photoelectric transmission usually requires the use of an opaque paper to reduce the amount of light transmitted directly through the paper stock, but there are no other restrictions on the color of the paper. For infrequent usage and the slow-speed handling of most applications, a relatively light and inexpensive paper stock can be used. However, if the paper tape is to receive any high-speed machine handling (e.g., 100 inches per second), or repeated use as a permanent tape in a control system, then it should be made from suitably durable stock. In special instances, plastic tape (such as Mylar tape) has been used for this purpose. Paper tape is usually used in a continuous reel form; however, unit-record card stock and continuous fan-fold tape are also used. The edge of a paper document (e.g., an 8½-by-11-inch page or a 3-by-5-inch card) could be punched in such a way as to permit the body of the paper to be used as a document, while the edge punching was used to carry some information for machine handling.

The basic families of paper tape codes are referred to as 5-, 6-, 7-, and 8-channel codes.[6] The "channel" refers to a coding line running along the length of a tape. A 5-channel tape, for example, has a pattern of 5 hole positions across the width of the tape. The hole combinations across the width of the tape determine the variety of codes possible. One row of hole positions across the width of a tape is called a "frame" and represents a single character, and is analogous to an IBM card column. Several codes are illustrated in Fig. 4-4. As with

[6] Tholstrup, H. L., "Perforated Storage Media," *Electrical Manufacturing*, Vol. 62, No. 6, pp. 53–61, 276 (December 1958).

Tholstrup, H. L., "Numerical Control by Punched Tape," *Instruments and Control Systems*, Vol. 34, No. 9, pp. 1643–1646 (September 1961).

Anon., "All About Paper Tape," *Datamation*, Vol. 5, No. 3, pp. 8–12 (May–June 1959); Vol. 5, No. 4, pp. 48–51 (July–August 1959).

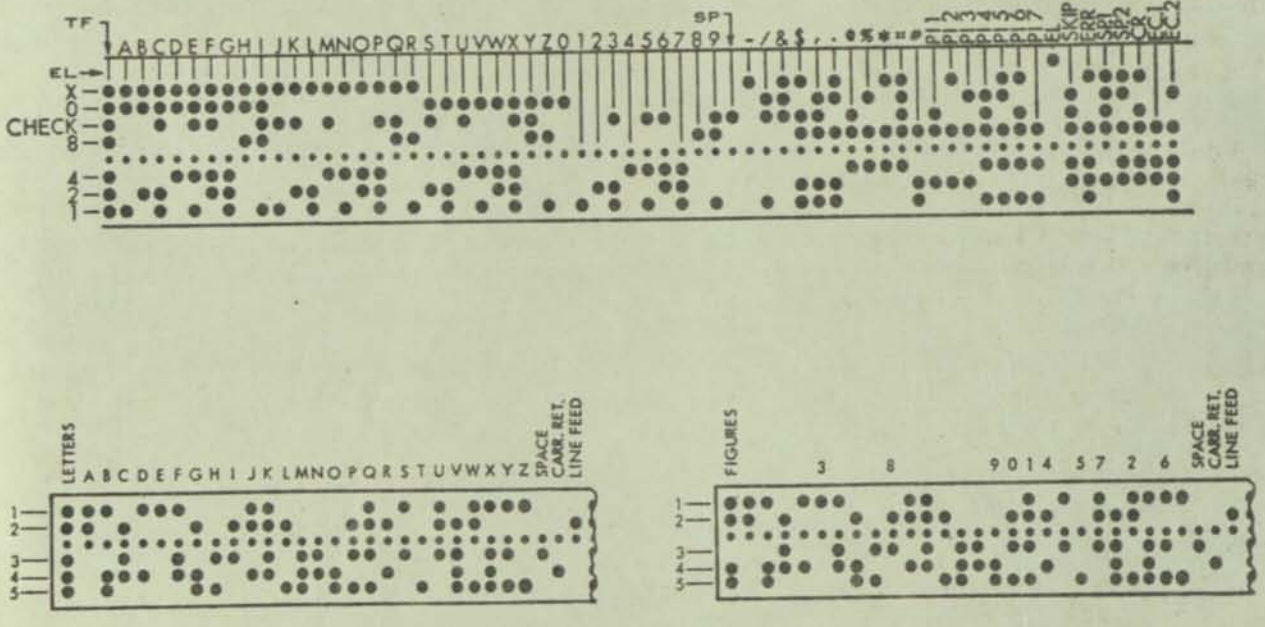


FIG. 4-4 Representative paper tape codes.

the tab cards, no single standard code exists, although several of the codes (notably the teletype and IBM codes) are used much more frequently than the others. The purchaser of the paper tape equipment usually has the final choice of codes. To handle code compatability problems, equipment is available to read a tape that is punched with one code and use this information to punch a tape in another code.[1] None of the common paper tape typewriter systems print symbols on the paper tape to correspond to the punched information, although the practice is followed by some teletype equipment.

C. MAGNETIC TAPE MEDIA

Information is stored on digital magnetic tape and magnetic cards by forming discrete-variable magnetic flux patterns in magnetic media according to some pre-determined code. The patterns are formed by passing an electric current through a recording head which has been placed very close to the surface of the magnetic medium. As the electric current passes through the recording head in a prescribed sequence, a corresponding sequence of magnetic field changes will occur. The magnetic material is held close enough to the recording head so that the magnetic material will be magnetized in the same pattern as the magnetic field of the original electric current. During recording, the magnetic medium is moved rapidly past the recording head so that new information is not written over old information. The materials used for the storage medium are such that they will permanently retain the flux patterns until the medium is remagnetized by another recording operation, or by some other magnetic field. As with any magnetic device, the medium will lose its magnetic patterns, and hence its memory, if it is subsequently exposed to high temperatures, extreme physical shocks, or large magnetic fields.

The magnetic pattern is normally invisible to the eye, so that recordings usually must be sensed electrically instead of visually. The sensing of the tapes is done with a reading head, which is similar to the writing head. When the storage medium is moved past the read head, the stored flux patterns of the medium induce a voltage in the head, which tends to follow the pattern of the stored flux. In this way, an electrical signal is obtained that corresponds to the recorded flux pattern. At this point, a clarification of the term "digital recording" would seem to be in order. Digital recording means that only discrete numbers or digits are recorded, most often with a binary number system. The information is represented by patterns of discrete-valued variables. The flux patterns for each coding position are usually oriented only in one of two allowed states, and numeric and alphabetic symbols are represented by uniquely specified groups of flux patterns. The "track" of the magnetic tape is analogous to the "channel" of the paper tape, and "frame" has the same meaning for both types of tape. Figure 4-5 shows a section of magnetic tape with 7 tracks or channels. Many magnetic tape systems have a 5-, 6-, or 7-channel code similar to that used with punched paper tape. In contrast, analog recording allows a continuous variation in the flux pattern in a manner that is analogous to the variations in the original stimulus. An audio tape recorder, for example, provides a recording in which the flux pattern is analogous to the original sound pattern. With a few noted exceptions, magnetic recording will, in this book, imply digital recording.

[1] Converter equipment is also available to read paper tapes and to transcribe the information onto other storage mediums, such as punched cards or magnetic tape. Conversely, equipment is available to transcribe information from many different mediums onto punched paper tape.

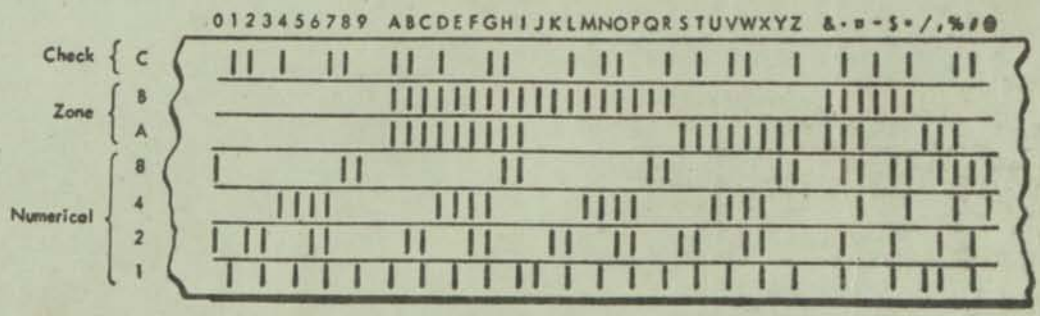


FIG. 4-5 Representative computer magnetic tape coding.

recognize that most of the limitations of each of the recording processes are intimately related to specific characteristics of the magnetic heads.

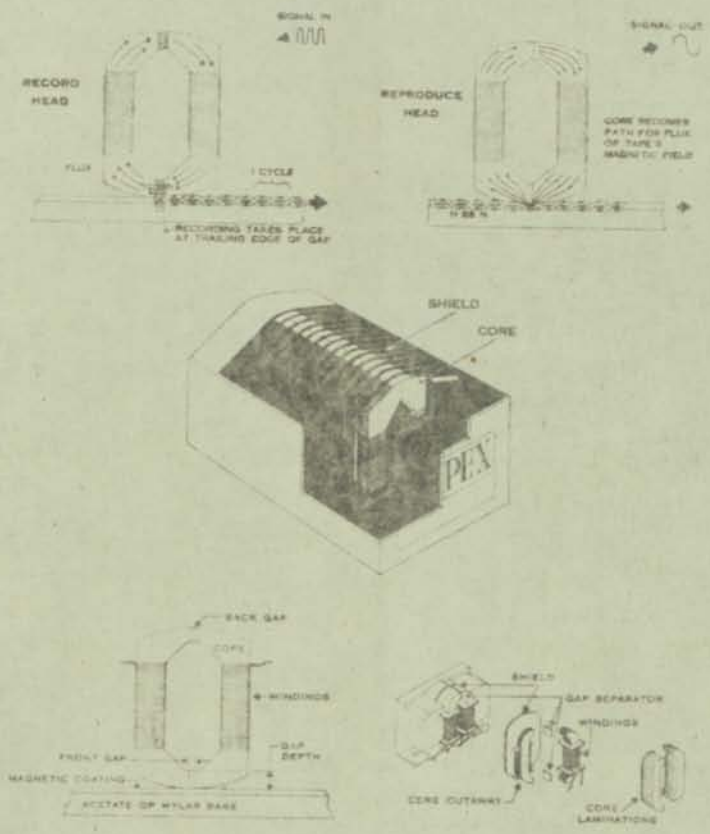


Fig. 26. Construction of a magnetic head.

Figure 26 illustrates the construction of a typical magnetic head. Two identical core halves are constructed of thin laminations of a material having high magnetic permeability, low electrical resist-

from Ampex publication "The Tape Recorder as an Instrument for Music" by Paul Weber, 1958

"Proposed USA Standard. Recorded Magnetic Tape for Information Interchange (200 CPI, NRZI)," CACM, Vol. 10, No. 11, pp. 730–739 (November 1967).

"Proposed USA Standard. Magnetic Tape Labels for Information Interchange," CACM, Vol. 10, No. 11, pp. 737–743 (November 1967).

A roll of magnetic tape is the usual storage medium, although magnetic cards have been used.[8] The magnetic tape normally consists of rolls of thin plastic (such as Mylar) ribbon that has been coated with a very thin layer of magnetic oxide material. In some cases, a plastic lamination is placed on top of the oxide (a "sandwich tape") to prevent the oxide from flaking or rubbing off during reading or writing. The rolls are normally 2400 to 3600 feet in length, and ½ inch, ¾ inch, or 1 inch in width. It is usually necessary to store and use the tape in a dust-free and controlled environment (nominally 50 to 85°F and 50 to 90% relative humidity) to prevent damage to the tapes or degradation of the information on the tape. If the tape is too dry, it tends to build up a charge of static electricity that may cause it to physically jam the tape transport mechanism. Combinations of moisture and high temperatures will cause the tape to stretch out of shape when it is being pulled at high speeds.

Since the information can be recorded in very small regions on the tape, it is possible to record several parallel tracks on the same tape. Figure 4-5 shows the coding and recording pattern of a representative digital magnetic tape. The information or pulse patterns can be stored very compactly on tape; typically there are from 5 to 20 tracks per tape, and from 100 to 1000 pulses per linear inch of track. The tape can also be moved at very high rates of speed, 75 to 200 inches per second. The high recording density and tape speed result in a rapid information transfer rate of 25 to 200 thousand digits or characters per second. And as with the punched card and paper tape systems, the physical dimensions, recording method, and coding system are parameters whose values have been established separately by almost every equipment manufacturer. Consequently, no single standard tape or coding system exists for digital magnetic tape. Only in a few instances can the magnetic tapes written on one manufacturer's computer be read on another manufacturer's computer.

MACHINE REPRESENTATION OF NATURAL LANGUAGES

Text Material

It would be most useful to have machines that could read ordinary English text, such as newspapers, books, and journal articles. Complex processing of text material—such as automatic indexing, automatic abstracting, machine translation, concordance construction, and automatic editing—cannot be done effectively or economically unless the processing equipment can store, recognize, and process the many different symbols and forms as used and recorded in our natural language. Present-day equipment is generally inadequate in this respect. Whereas the text of a published dictionary may currently use an alphabet of approximately 1400 different characters (uppercase and lowercase letters of 4 different type styles plus special characters), most data processing equipment is only designed for a total alphabet size of something less than 90 characters. For this reason, many coding tricks and simplifications must be used to represent text material in machine language today. Several schemes have already been proposed to represent text material in the language of the current data processing machinery.[9]

Image Material

Chemical structures (such as molecules and polymers) and their symbolic notation or coding are a part of our natural language that is usually difficult to represent in machine language. Several coding procedures have been developed: [10] semi-automatic

[8] The magnetic cards will be discussed later (Chapt. 8) in conjunction with the Magnacard system.

[9] Nugent, W. R., "A Machine Language for Document Transliteration," paper presented at the 14th National Conference of the Association for Computing Machinery, Massachusetts Institute of Technology (September 1959).

Newman, S. M., R. W. Swanson, and K. C. Knowlton, "A Notation System for Transliterating Technical and Scientific Texts for Use in Data Processing Systems," Chapt. 7 in *Information Retrieval and Machine Translation, Part 1* (Interscience Publishers, New York, 1960).

[10] Waldo, W. H., and M. DeBacker, "Printing Chemical Structures Electronically: Encoded Compounds Searched Generically with IBM-702," *Proceedings of the International Conference on Scientific Information*, Vol. 1, pp. 711–730 (National Academy of Sciences, Washington, D.C., 1959).

Opler, A., and N. Baird, "Display of Chemical Structure Formulas as Digital Computer Output," *American Documentation*, Vol. 10, No. 1, pp. 59–63 (January 1959).

Opler, A., and N. Baird, "On the Automatic Manipulation of Representations of Chemical Structures," *American Documentation*, Vol. 10, No. 2, pp. 130–134 (April 1959).

Opler, A., "A Topological Application of Computing Machines," *Proceedings of the 1956 Western Joint Computer Conference*, pp. 86–88.

Wheeler, K. W., "A Structure Code for Organic Compounds," *American Documentation*, Vol. 9, No. 3, pp. 198–207 (July 1958).

The information recorded on digital tape is seldom represented as one continuous character string. For a number of reasons, the recording is usually made in smaller blocks of records. The blocks may be of the same fixed length (e.g., 1000 characters) for all records on the tape, or ~~it~~ the block length might vary with the size of each record (e.g., a separate tape record for each library catalog card record). The size of the blocking depends on the particular application, equipment and programs available. The record boundaries are identified on the tape, so that it is possible for the programmer to issue computer instructions such as "Read 3 records." Even with the spaces of blank tape serving as boundaries between the individual records, it is still possible to record million ? characters on a single roll of tape. An indication of the storage capacities available with various magnetic medias is given in Fig. .

Magnetic cards used with some computer systems operate on the same basic principle as magnetic tape, but usually have a format and coding system unique to that manufacturers' equipment. Examples of such magnetic card systems are the NCR CRAM, the RCA ? , and the IBM Data Cell.

Illustrations of several types of magnetic cards and their cartridges are given in Fig. .

In addition to these card systems, magnetic ledger forms have also been developed for use with some accounting machines. These ledger forms are regular paper stock as used with many bookkeeping machines, but with an oxide material coated in strips on the back of the page to record the typed information, or portion thereof (e.g., current balance) in a machine language form.

Magnetic tape cartridges are also used with some electric typewriters as a magnetic equivalent of a paper tape typewriter. These systems are described later in Chapter .

Magnetic drums, and magnetic discs are also used as a means of computer storage, and their formats and coding systems are also usually established independently by each manufacturer. Drums can be thought of as continuous rings of tape wrapped around a drum ranging in size from the bulk of an electric fan motor to that of a 50 gallon oil drum. Magnetic discs store information on annular rings, and range in size from a small phonograph record to a diameter of 3 - 4 feet.

Another form of the magnetic media is that used for the internal storage of a computer. The word internal usually refers in this case to the fastest memory circuitry or equipment ~~that is~~ operating as an integral part of the computer, and is sometimes also referred to as the working or high speed storage. In most modern computers, this internal storage is achieved with magnetic core or thin magnetic film components. Their method of storage can be thought of as somewhat analogous to magnetic tape recording -- that is, ~~showing~~ representing a binary state by having a core or local portion of a film magnetized in one of two different states. This internal storage is usually represented, as a total ~~allocation~~ collection of smaller modules, with modules defined for various machines as digits, characters, bytes, and words. One manufacturers' computer may be described using one definition for memory size (e.g., 32,000 words of main memory), while another manufacturer uses his definition for his machines' memory (e.g, 64K bytes of main memory). The reader can resolve these differences by referring to the following brief table of definitions:

bit = binary digit, the most elemental number, having the value equal to zero or one.

digit = one digit, an ambiguous term unless qualified by describing the number base (e.g., binary digit, octal digit, decimal digit)

character= one character, an ambiguous term unless qualified by describing the number of bits that are used to describe that character (e.g., a 5-bit character, an 8-bit character)

byte = an expression used primarily by IBM to mean a set of 8 bits that normally represent a single alphabetic or special character, or two decimal digits.

word = one word, an ambiguous term unless qualified by identifying the particular computer, because different computer models often have different word sizes (e.g., 16-bit CDC-160 word, 64-bit IBM 7094 word).

An illustration of several types of memory circuits is given in Fig. .

With increased use of computers for composition and printing purposes, more consideration was given to ways of increasing the available character set. Some of the IBM computers made provision for storing and printing both upper and lower case characters as well as a number of special characters. This permitted a number of libraries to print catalog cards or book catalogs at high speed in upper and lower case form. Examples of some of this printing is given in Figs . An illustration of the character set used with the high speed printer for several research libraries is given in Fig. . The use of these large alphabets presents some difficulties with the predominantly single-case equipment used to convert this information into machine language form. An example of one code used to represent both upper and lower case information in punched card form is given in Table . This code was used at Harvard University during the conversion of the Widener shelf list.* (*Ref.) An example of one code used to represent both upper and lower case information in paper tape form is given in Table . This paper tape code is the one used in the pilot MARC project at the Library of Congress, and represents the character set described as the Standard Library Keyboard* (Ref.).

Other paper tape codes have been developed to represent the large number of characters used in the typesetting of a scientific journal. The American Chemical Society for example uses a paper tape code to represent up to 1440 different characters, using a standard paper tape typewriter ~~key~~ keyboard,* ~~(*Ref., Kinney,~~

Davenport, W.C. and J. T. Dishaw, "Computer-Based Composition at Chemical Abstracts Service," JCHD, Vol. 6, No. 4, pp. 221-225 (November 1966).

Kinney, J. H., et al., "Computer-Aided Typesetting for the Journal of Chemical Documentation", JCHD, Vol. 6, No. 1, pp. 1-2 (February 1966).

Paper tape typewriters have also been developed to handle representation of chemical structures, including punching and reading of paper tape, & the typing of appropriate symbols.*

Feldman, A. et. al., "An Automatic Encoding of Chemical Structures," JCHD, Vol. 3, No. 4, pp. 187-189 (October 1963). Walter Reed Chemical typewriter.

Mullen, J. M., "Atom-by-Atom Typewriter Input for Computerized Storage & Retrieval of Chemical Structures," JCHD, Vol. 7, No. 2, pp. 88-93 (May 1967). Shell typewriter.

5

D. Character Readers

Character reading equipment represents a more recent development in input equipment that has continued to receive increasing use, even though there are still a number of constraints in using this equipment in order to obtain the necessary accuracy and quality control of the reading process. The following sections describe several types of character readers. Excellent summaries and reviews of this field have been prepared by Falk.* (*Ref. Falk, Howard, "Optical Character Recognition Systems," Electro-Technology, pp. 42, July 1964) and by .

1. Handwritten Characters

After many years of trial and testing, character recognition equipment is in operational use today that will read numbers that are handwritten in a partially constrained style. This equipment is being used primarily for reading such things as retail sales slips, telephone toll tickets, and other forms that require a brief numeric entry. An example of the type of characters that can be recognized is given in Fig. . Several models of equipment are available, including the IBM 1287 and the Rabinow illustrated in Figs. and , respectively.

2. Page Readers

A large amount of character recognition equipment is in use today in the banking industry to read single lines of information printed in magnetic ink on checks and check-size documents. Some equipment is being used to read the ZIP codes on mail envelopes. Some equipment is also in use to read the information from rolls of adding or cash register tape (journal tapes). But the most interesting equipment from the point of view of library operation is the equipment to read many lines of printed information from a page at relatively high speeds.

The page reading equipment is available in two general versions: a minimum system operating with a single stylized type style; and a more flexible and expensive system that can read any of several type styles.

Several type styles are in use today, including the USASI Standard Optical Character Recognition Font * (*"Proposed American Standard: Character Set for Optical Character Recognition." CACM, Vol. 8, No. 1, pp. 18-26 (January 1965), the IBM 1428 font, the National Optical font (NOF) developed by the National Cash Register Co., and the Farrington Selfcheck font, both of which are illustrated in Fig (7B) . These fonts are highly stylized in order to ensure the uniqueness or seperability of the characters from the machine's point of view, while at the same time providing acceptable human legibility (which you have not provided you secretary in your draft!). This enhances the reject and accuracy performance of the equipment while holding down the costs. In order to provide relatively error-free performance, these characters must be typed under good quality control procedures with typewriters which are equiped with the appropriate character sets. In the library field, this equipment has been used for the conversion of catalog data to machine language form. An illustration of a representative input form is given in Fig. . Acceptable page sizes generally range from 3 x 5 inch cards to legal size paper sheets, ~~Examples of some of the available single-font page reading equipment are the CDC 915, the Farrington.~~ This equipment operates in the general range of 350 characters per second.

Equipment is also available to recognize and interpret several different type fonts. The IBM 1287 for example, in addition to reading handwritten numerals, also reads the IBM 1428, ASCS, NOF, and 7B fonts. A special page reader developed for the Social Security Administration can read 200 different type faces at rates up to 650 lines per minute. No equipment is available yet, or likely to be generally available for some time, that can read conventional printed matter such as library catalog cards, or printed books or magazines, with any satisfying cost or performance. Further information on page reader techniques is given in Chapter .

Rule	Correct	Incorrect
1. Write big.	01834	0 2 8 3 4
2. Close loops.	06889	06889
3. Use simple shapes.	02375	02375
4. Do not link characters.	00881	00887
5. Connect lines.	45T	45T
6. Block print.	CSTXZ	CSTXZ

Figure 6. Handwriting Rules

1287 Standard Fonts		NOF* (Special Feature)	78** (Special Feature)	Data Byte Configurations							
IBM 1428 Font	ASCSOCR			0	1	2	3	4	5	6	7
0	0	0	0	1	1	1	1	0	0	0	0
1	1	1	1	1	1	1	1	0	0	0	1
2	2	2	2	1	1	1	1	0	0	1	0
3	3	3	3	1	1	1	1	0	0	1	1
4	4	4	4	1	1	1	1	0	1	0	0
5	5	5	5	1	1	1	1	0	1	0	1
6	6	6	6	1	1	1	1	0	1	1	0
7	7	7	7	1	1	1	1	0	1	1	1
8	8	8	8	1	1	1	1	1	0	0	0
9	9	9	9	1	1	1	1	1	0	0	1
C	#	ô		1	1	0	0	0	0	1	1
N	@	₱		1	1	0	1	0	1	0	1
S		₫		1	1	1	0	0	0	1	0
T	¥	¥		1	1	1	0	0	0	1	1
X		₩		1	1	1	0	0	1	1	1
Z		₡		1	1	1	0	1	0	0	1
≠ or |***	|***			0	1	1	0	0	0	0	0
blank	blank	blank		0	1	0	0	0	0	0	0
reject (⊟)	reject (⊟)	reject (⊟)		0	1	1	1	1	1	0	0
special blank	special blank			0	1	1	1	0	0	0	0

* National Optical Font shown by permission of National Cash Register Co.
** 78 Font shown by permission of Farrington Manufacturing Company.
*** These symbols are used as field separators.

Figure 1. Fonts

from IBM A28-
A21-

- THE TYPE FONT -

LETTERS

A B C D E F

G H I J K L

M N O P Q R

S T U V W X

Y Z

NUMBERS

1 2 3 4 5

6 7 8 9 0

STANDARD SYMBOLS

. PERIOD
, COMMA
? QUESTION MARK
/ SLANT OR SLASH
' APOSTROPHE
" QUOTATION MARK
- HYPHEN
(LEFT PARENTHESES
) RIGHT PARENTHESES
: COLON
; SEMI-COLON
* ASTERISK
= EQUAL SIGN
+ PLUS SIGN
% PERCENT SIGN
$ DOLLAR SIGN
& AMPERSAND
# NUMBER SIGN

ABSTRACT SPECIAL SYMBOLS

{PROGRAMMED USE}

¢	CANCEL	↓	ARROW DOWN
⅄	FORK		
Δ	DELTA	↑	ARROW UP

UNIQUE INFORMATION SYMBOLS

THE READER WILL RESPOND TO THESE UNIQUE SYMBOLS AS DESCRIBED.

—— STRIKEOUT. A HORIZONTAL LINE MORE THAN 2 AND 1/2 CHARACTER SPACES IN LENGTH { —— }. THE LINE MAY BE TYPED OR DRAWN IN.

| FIELD SEPARATOR. A VERTICAL LINE OF MORE THAN 1 AND 1/2 CHARACTER HEIGHTS. READER CAN RECOGNIZE IT AS A "COLUMN" INDICATOR. NORMALLY, SUCH A LINE WILL BE PRE-PRINTED ON STANDARD FORMATS PREPARED FOR USE WITH THE READER.

■ CIRCLE FILLED.
□ CIRCLE UNFILLED. THE FILLED AND UNFILLED CIRCLES ALLOW FOR THE MANUAL ENTRY OF DATA UTILIZING THE "MARK SENSE" MODE. THE UNFILLED CIRCLE IS NORMALLY PRE-PRINTED. THE READER WILL DETECT WHETHER IT HAS BEEN FILLED OR NOT, AND RESPOND IN ACCORDANCE WITH THE PROGRAM.

NOTE:

IF UNIQUE RECOGNITION OF THE SEMI - COLON, APOSTROPHE, OR COMMA IS A MANDATORY REQUIREMENT, THE FOLLOWING CHARACTER SHAPES ARE RECOMMENDED FOR THE FONT:

SEMI-COLON APOSTROPHE COMMA

from CDC brochure 40811600

simulation of character and pattern recognition logic; (2) automatic encoding of chemical structures, electrical circuits, or other line drawings; (3) object counting, such as particle or bacterial colony counts; (4) determination of areas of selected terrain or vegetation on aerial photographs; (5) analysis of tracking data such as photos of satellite tracks against a star background; (6) analysis of astronomy photographs to compute positions and motions, and to establish star catalogs; (7) generation of elevation contour lines from aerial photographs; (8) selection of optimum road routing and cut-and-fill computations from contour maps; (9) superimposition of images to form composite images; (10) improvement of contrast on such low-definition images as X-ray photographs, and the enhancement of these images to accentuate outlines and skeletal features. This is already being done for the pictures obtained by some U.S. space satellites. It seems reasonable to expect that computer equipment will approach most of these other capabilities with some success in the next 5 or 10 years.

F. ASCII Code (insert pg 79a)

Additional References

Bemer, R. W., "Survey of Coded Character Representation," *Communications of the Association for Computing Machinery*, Vol. 3, No. 12, pp. 639–642 (December 1960).

Bonnett, H. T., and D. W. Calhoun, "Application of Line Formula Notation in an Index of Chemical Structures," *Journal of Chemical Documentation*, Vol. 2, No. 1, pp. 2–6 (January 1962).

Doyle, W., "Recognition of Sloppy, Hand-Printed Characters," *Proceedings of the 1960 Western Joint Computer Conference*, pp. 133–142 (Institute of Radio Engineers, New York).

Frishkopf, L. S., and L. D. Harmon, "Machine Reading of Cursive Script," Chapt. 23 in *Information Theory*, C. Cherry, editor (Butterworth, London, 1961).

Garfield, E., "Information Theory and Other Quantitative Factors in Code Design for Document Card Systems," *Journal of Chemical Documentation*, Vol. 1, No. 1, pp. 70–75 (January 1961).

Gull, C. D., "Holes, Punches, Notches, Slots, and Logic," Chapt. 19 in *Punched Cards*, 2nd ed., R. S. Casey et al., editors (Reinhold Publishing Corp., New York, 1958).

Hayward, H. W., *A New Sequential Enumeration and Line Formula Notation System for Organic Compounds*, U.S. Patent Office, Washington, D.C., Research and Development Report 21 (November 1961).

Kirsch, R. A., et al., "Experiments in Processing Pictorial Information With a Digital Computer," *Proceedings 1957 Eastern Joint Computer Conference*, pp. 221–229 (Institute of Radio Engineers, New York).

Neisser, U., and P. Weene, "A Note on Human Recognition of Hand-Printed Characters," *Information and Control*, Vol. 3, No. 2, pp. 191–196 (June 1960). See also AD-236 395.

Oliver, B. M., "Efficient Coding," *Bell System Technical Journal*, Vol. 31, No. 4, pp. 724–750 (July 1952).

Rosen, L., "Characteristics of Digital Codes," *Control Engineering*, December 1959, pp. 115–119.

Ross, H. M., "Considerations in Choosing a Character Code for Computers and Punched Tapes," *Computer Journal* (British), Vol. 3, No. 4, pp. 202–210 (January 1961).

Ross, H. M., "Further Survey of Punched Card Codes," *Communication of the Association for Computing Machinery*, Vol. 4, No. 4, pp. 182–183 (April 1961).

Smith, H. J., and F. A. Williams, "Survey of Punched Card Codes," *Communications of the Association for Computing Machinery*, Vol. 3, No. 12, p. 638 (December 1960).

Starker, L. N., and J. A. Cordero, "The Cyanamid Organic Structure Code and Search System," *Journal of Chemical Documentation*, Vol. 2, No. 1, pp. 12–15 (January 1962).

Waldo, W. H., "Searching Two-Dimensional Structures by Computer," *Journal of Chemical Documentation*, Vol. 2, No. 1, pp. 1–2 (January 1962).

Hoelskins, J.A., "Paper Tape vs Punched Cards,"

Feldman, A., "A Proposed Improvement in the Printing of Chemical Structures, which Results in their Complete Computer Codes," AMDO, Vol. 15, No. 3, pp. 205–209 July 1964

F. ASCII Code

The American Standard Code for Information Interchange (ASCII) was approved in June 1963, and is shown in Fig. . This code, meant to be used for communication purposes between various data processing equipment, is a 7-level code with 128 code combinations. Of these, 64 are assigned to graphic symbols, 36 to control codes, and 28 are reserved for future use. Several subsets are possible with this code, including a 64-character subset for 6-bit applications, and a 4-bit subset for numeric applications. Some standards work is also being done to lead to implementation of this code in punchard cards, paper tape, and magnetic tape.*

*Brookman, R. G., "Standards" Data Systems Design pp. 15-18 (November 1964).

b_7				0	0	0	0	1	1	1	1
b_6				0	0	1	1	0	0	1	1
b_5				0	1	0	1	0	1	0	1
b_4	b_3	b_2	b_1								
0	0	0	0	NULL	DC_0	b	0	@	P	UNASSIGNED	ACK
0	0	0	1	SOM	DC_1	!	1	A	Q		
0	0	1	0	EOA	DC_2	"	2	B	R		①
0	0	1	1	EOM	DC_3	#	3	C	S		ESC
0	1	0	0	EOT	DC_4 (STOP)	$	4	D	T		DEL
0	1	0	1	WRU	ERR	%	5	E	U		
0	1	1	0	RU	SYNC	&	6	F	V		
0	1	1	1	BELL	LEM	(APOS)	7	G	W		
1	0	0	0	FE_0	S_0	(	8	H	X		
1	0	0	1	HT/SK	S_1	)	9	I	Y		
1	0	1	0	LF	S_2	*	:	J	Z		
1	0	1	1	V_{TAB}	S_3	+	;	K	[		
1	1	0	0	FF	S_4	(COMMA)	<	L	\		
1	1	0	1	CR	S_5	−	=	M	]		
1	1	1	0	SO	S_6	.	>	N	↑		
1	1	1	1	SI	S_7	/	?	O	←		

UNASSIGNED CONTROL

EOM	End of message
EOT	End of transmission
WRU	"Who are you?"
RU	"Are you . . .?"
BELL	Audible signal
FE_0	Format effector
HT	Horizontal tabulation
SK	Skip (punched card)
LF	Line feed
V_{TAB}	Vertical tabulation
FF	Form feed
CR	Carriage return
SO	Shift out
SI	Shift in
DC_0	Device control reserved
NULL	Null/Idle
SOM	Start of message
EOA	End of address
DC_1-DC_3	Device control
DC_4(Stop)	Device control (stop)
ERR	Error
SYNC	Synchronous idle
LEM	Logical end of media
S_0-S_7	Separator (information)
b	Word separator (space, normally non-printing)
<	Less than
>	Greater than
↑	Up arrow (Exponentiation)
←	Left arrow (Implies/ Replaced by)
\	Reverse slant
ACK	Acknowledge
①	Unassigned control
ESC	Escape
DEL	Delete/Idle

FIG. 1 American Standard Code for Information Interchange.

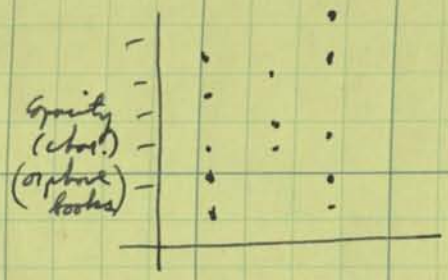
element; 2, it would appear justifiable to include the transmission of a character parity bit; and 3, there are media, systems and equipment considerations that argue for even character parity. It was also concluded that consistent character parity will facilitate systems maintenance and operation, and will simplify equipment in multiple-line systems.

ASCII. The American Standard Code for Information Interchange was approved June 17, 1963. The code is shown in Fig. 1. The abbreviations in the code have the following meanings:

The ASCII is a 7 level code with 128 code combinations. Of these, 64 combinations are assigned to graphic symbols, and the remainder are assigned to controls. In addition to these two assigned areas a portion of the code set has been reserved for future use. The graphic area of the code is a dense subset containing the alphabet, numbers, punctuation marks and special symbols; there are no controls within this subset. The control area of the code consists of controls that logically fall into four groups: 1, transmission controls; 2, format effectors; 3, device controls; and 4, information separators. Use of these controls is discussed in more detail later. One of the major advantages the ASCII controls have over the Baudot controls is that basic data communication control can be implemented with single character representations.

There are several subsets possible with ASCII including a 64-character subset for 6-bit applications, and a 4-bit subset for numerics only applications.

Subcommittee X3.2 is now engaged in work which will lead to the implementation of ASCII in punched cards, paper tape and magnetic tape. In addition to these efforts, work is taking place in Subcommittee X3.3 on line control techniques, character structure (including bit sequencing) and signaling speeds. Both subcommittees are engaged in developing better definitions for some codes in the standard.



Equip.			Storage Capacity
IBM 1302 Mod 1	disc		117 mill. char.
IBM 1302 Mod 2			234 " "
IBM 1301 Mod 1			28 " "
IBM 1301 Mod 2	↓		56 " "
IBM 2311		disk	7.25 mill bytes per disk pack
IBM 2302 disk		Mod 3	113 mill. bytes or 226 mill packed decimal digits
↓		Mod 4	226 " 452 " "
IBM 2321 Data Cell Drive			over 418 mill. bytes per 2321.
IBM 2303 Drum Storage			4 mill. bytes or 8 mill. packed dec. digits
IBM Hypertape			~~55~~ mill. bytes per reel

Weber, P. "The Tape Recorder as an Instrumentation Device." August 1958

creates a weight, size, and complexity problem under certain conditions, such as when recording data in flight.

Applications of the Digital Recording Process

Let us summarize the *advantages* of the Digital Recording Process:

1. Inherent capability of extremely high orders of accuracies.
2. Recording relatively insensitive to tape transport speed instabilities.
3. Simple record and reproduce electronic circuitry.
4. Output information is in the proper form for feeding directly into digital computers.

The relative *disadvantages* of the Digital Recording Process can be summarized as follows:

1. Poor tape economy – 1/8th that of the PDM process and 1/30th that of the FM process.
2. Data must be digitized at the source, or special digital transducers must be employed.
3. Reliability extremely dependent on tape quality, requiring redundancy and parity-check features.

The applications of digital recording are primarily for the processing of edited data, involving digital-computer techniques; and as an input device, output device, and internal storage for digital computers.

MAGNETIC RECORD AND REPRODUCE HEADS

We are now ready to discuss the second basic element of a magnetic tape recorder—the *Magnetic Head*. It is no mere copywriter's slogan to refer to the magnetic head as the heart of a recorder. To manufacture a series of practical magnetic recorders, which can be universally used for instrumentation purposes, requires highly refined and precise production and quality-control techniques. The objective must be to produce heads having absolute uniformity of electrical and mechanical properties, to permit multiple track recordings to be interchanged from machine to machine without mechanical alignment or adjustment; to permit amplifiers to be substituted one for the other, and connected interchangeably from track to track without matching or tuning; and to permit any track to be used with any of the other four basic recording processes. And recalling the previous theoretical discussions, we

recognize that most of the limitations of each of the recording processes are intimately related to specific characteristics of the magnetic heads.

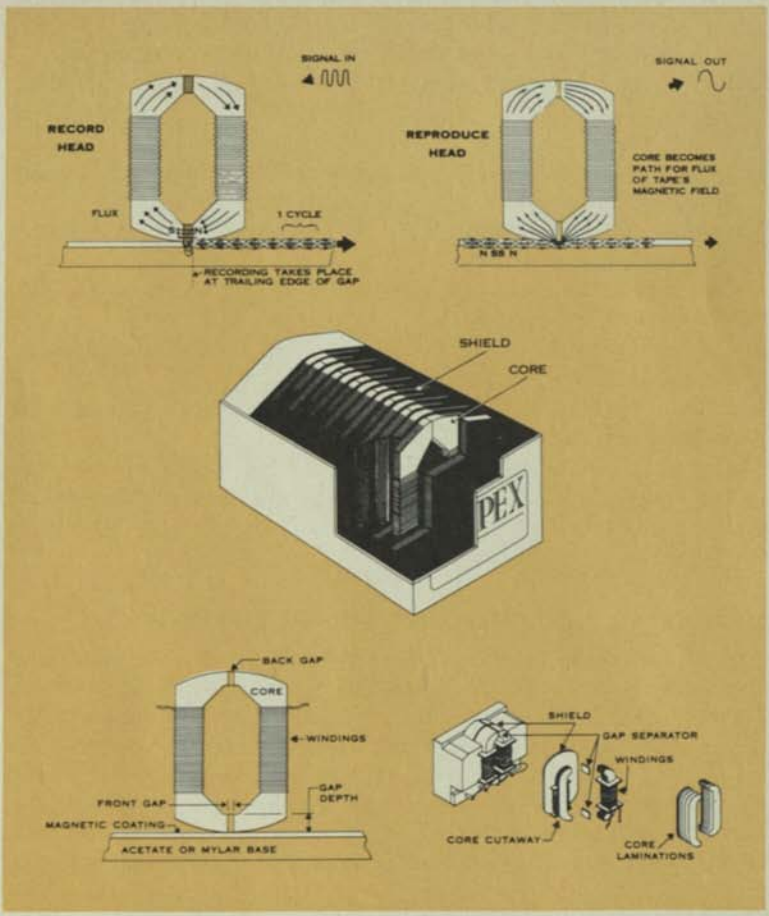


Fig. 26. Construction of a magnetic head.

Figure 26 illustrates the construction of a typical magnetic head. Two identical core halves are constructed of thin laminations of a material having high magnetic permeability, low electrical resist-

Appendix
4-Mode
Table of Random n Codes for Superimposed Code Systems with 40-position Fields

Appendix

Table of CODENS used in this Book