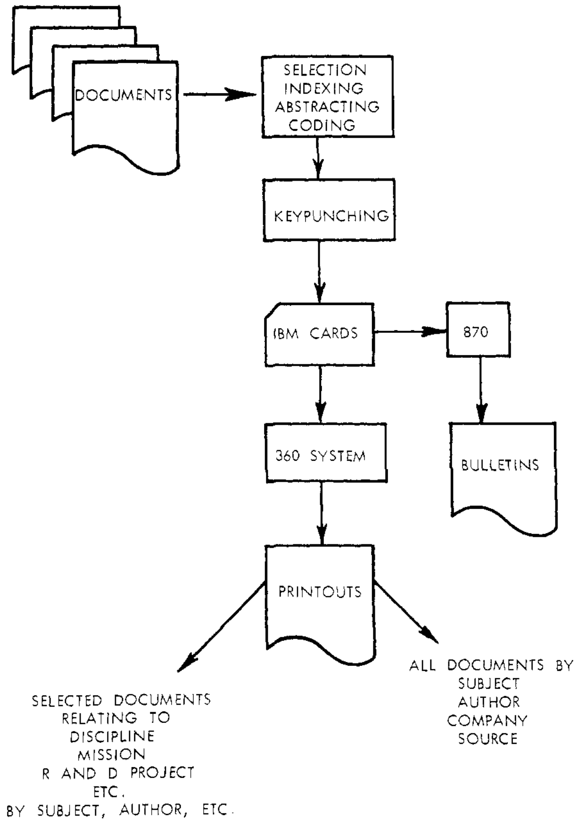


Figure 4. Flowchart of information system

differences among the members and the over-all objectives of a community of scientists constitute an array of variables, such as disciplines and missions of science, R and D projects, and individual and group interests. Within these differences and objectives, the information needs of the scientists are resolved by two mechanisms: current awareness bulletins and retrieval systems.

Traditionally, these two mechanisms, awareness and retrieval, have been separate operations (*1, 5*); also, documents processed from different viewpoints have been handled as many times as there were viewpoints, or alternatively, the documents have been indexed and abstracted from the viewpoint of the documents without relationship to the needs and requirements of the community of users.

The uniqueness of the information system described in this paper is the mechanized flow from a single input to multiple information products as shown in Figure 4. This mechanized flow has been made possible by the combining of an IBM 870 Document Writing System with the IBM System/360. The objective of this paper has been to treat the information system from the perspective of its uniqueness. Subsequent papers will detail the operating aspects of the machines and of the information system for specific needs and requirements of the community of scientists it serves.

LITERATURE CITED

(1) Friedenstein, H., "Alerting with Internal Abstract Bulletins," J. CHEM. DOC. **5**, 154-7 (1965).
(2) Skolnik, H., "The Hercules Literature Chemist," *Hercules Chemist*, No. 41, 7-9 (February 1961).
(3) Skolnik, H., Chap. 7 in "Vistas in Information Handling," Vol. 1, edited by P. W. Howerton and D. C. Weeks, Spartan Books, 1963.
(4) Sorrows, H. E., "Industrial Technical Intelligence," *Research Management* **10**, 217-27 (1967).
(5) Strauss, L. J., I. M. Strieby, and A. L. Brown, "Scientific and Technical Libraries," Chap. 10, Interscience, 1964.
(6) "CAS Today," Chemical Abstracts Service, 1967.
(7) "IBM System/360 Principles of Operation," 6th ed., IBM.
(8) "Reference Manual. IBM 870 Document Writing System," IBM (November 1961).

Error Control in a Computerized Coordinate Index/Document Retrieval System*

J. L. HOLLOWELL†
Marshall Laboratory, F. & F. Dept., E. I. du Pont de Nemours, Philadelphia, Pa.

Received December 8, 1967

A novel technique has been developed for making substantial reductions in indexer, clerical, and keypunch-derived errors. In use for over 2 years in a medium-sized document retrieval system, real benefits included shortening of "clean-up" time after computer up-dates, less noise in the system for surer searches, and shortened keypunch time. The retrieval system comprises a term coded thesaurus with automatic generic posting, a term-document search dictionary with extensive link and role usage, and a doc-term file. Both machine and manual searches are made.

Error control in a computerized information and document retrieval system is often a significant and onerous problem. This paper describes a group of techniques, some novel, some otherwise, which have been used successfully for the past 2½ years to make substantial reductions in indexer, clerical, and keypunch-derived errors. These techniques have also simplified and speeded up several of the basic processes for inputting and processing information in the system, especially shortened keypunch time, shortened post-update "clean-up," and reduced "noise" in the system.

These techniques of error control have been applied

*Presented in Symposium on "Automation of Information Operations," ACS Middle Atlantic Regional Meeting, Philadelphia, Pa., February 1, 1968.
†Present address, D-7012, F. & F. Dept., E. I. du Pont de Nemours & Co., Inc., Wilmington, Del. 19898

in a coordinate indexing scheme using links, roles, and automatic post-up with provisions for both manual and machine searching. Computer printouts, updated about three times a year, provide the primary tools: a thesaurus of 13,000 chemical and technical terms and an inverted file of 75,000 term/document entries. This system serves a group of approximately 200 research and development people in an industrial coating research laboratory and includes documents of a wide range of complexity, from sales memoranda to research reports and patents.

Fundamental to this system is the use of a seven-character code for each term in the thesaurus; a typical term entry being:

1059000 EPOXY ESTER RESIN

It is the use (or misuse) of these codes around which the present subject of error control is built.

Figure 1 illustrates our basic input and update operation.

This system was initially operated for about one year when it became obvious that the amount of error in the system was consuming an inordinate number of man hours for detection and correction at each updating and could not be further tolerated.

Errors of the following type were found (in the approximate order of importance):

> Writing, memory, and recognition errors of the term code at indexing and at term code assignment.
>
> Keypunch errors: displaced fields, misread digits, transpositions, faulty verification, trash cards, machine stuttering.

These errors were in turn often compounded because invalid cards at the card-to-tape step led to aborted input sequence at the update. Substantial turnover of personnel has been a contributing factor as well.

The consequences of such errors were, of course, confusion in a) the thesaurus—missing terms and missing or inappropiate relationships, b) the inverted file—incomplete searches and unnecessary document screening because references were either missing or misplaced.

Such recitation is presumably all too familiar to those who have operated similar computerized systems. At one point, nearly 10% of our input of terms and term/document/role entries for a given updating was suspect and required review. This meant finding the error, diagnosing the trouble, preparing deletions of erroneous entries and preparing correct re-input of the initial information.

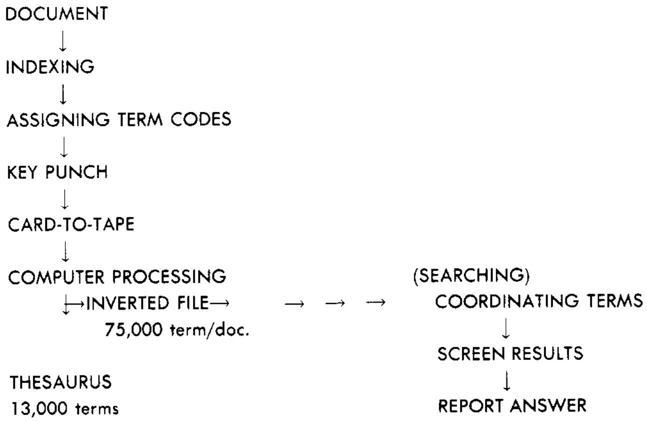


Figure 1. Flow chart of basic input operation

SOLUTIONS TO ERROR PROBLEM

The following techniques have been adopted to bring error incidence under control: check digits, rubber stamp thesaurus, hole-in-matrix, listing, automatic document listing, mnemonic term codes, and internal checks.

Transcription and keypunching of the term codes were our prime sources of error. Accordingly, conversion to a check digit basis was undertaken. The use of check digits for codes is a well-known scheme in inventory, banking, and related business system control but seems to be poorly recognized in document retrieval systems. In it, one of the digits of an alpha-numeric code is designated to serve as a verifier for a calculation based on the remaining digits (there are several publications by IBM, and others on these calculations. See IBM Bulletin Series **1200**, "Account Numbering and Self-Checking Number Systems"). A typical term code appears:

G659004 EPOXY ESTER RESIN

We use a routine for check digit calculation related to that designated by IBM as Modulus 10. The routine is shown in Figure 2.

Basically every other digit of the code is multiplied by 2 and then the sum of the final individual digits is obtained. This sum, 26, is then subtracted from the next higher multiple of 10, i.e. 30. The remainder, 4, is the check digit. Note that the sum is of the *individual* digits, not of the products as we would conventionally think.

The values assigned to the letter in the code are shown in the following chart (Figure 3). These are derived simply by summing the conventional digit-punch values with the values for the zone-punches: zone punch 12 has a value of 10, zone punch 11 is 20, and zone punch 0 is 30.Note that this latter operation in reality adds respectively 2, 4, or 6 to the final digit summation (see Figure 2).

Conversion of our established system to check digits posed several complications. Our initial seven-digit term codes—e.g., 1059000 EPOXY ESTER RESIN—gave a maximum potential of 10 million terms in the thesaurus. Loss of one digit for checking would cut the potential one tenth, to 1 million. This was below a practical mini-

```
                                          check
            G     6     5     9     0     0     —
            7
           10
           ──
           17
multiply    2     2           2           2
           ──────────────────────────────────────
           27    12     5    18     0     0     —
summing     2
            7
            1
            2
            5                     30 −26 = 4
            1
            8
            0
            0
           ──
           26
```

Figure 2. Check digit calculation

	DIGIT PUNCH	ZONE PUNCH	ZONE VALUE USED	SUM	NET CONTRIBUTION TO CHECK DIGIT
A	1	12	10	11	3
B	2			12	4
C	3			13	5
D	4			14	6
E	5			15	7
F	6			16	8
G	7			17	9
H	8			18	0
I	9			19	1
J	1	11	20	21	5
K	2			22	6
L	3			23	7
M	4			24	8
N	5			25	9
O	—			—	—
P	7			27	1
Q	8			28	2
R	9			29	3
S	2	0	30	32	8
T	3			33	9
U	4			34	0
V	5			35	1
W	6			36	2
X	7			37	3
Y	8			38	4
Z	9			39	5

Figure 3. Letter values for check digits

mum because a number of term areas in the thesaurus were already crowded. On the other hand, the computer programs were restricted to a seven-digit field, and the alternative of rewriting to encompass eight digits was impractically expensive. We hit upon a happy compromise: we converted the first digit to a letter (excluding the letter "O") and were able to maintain a 2.5-million term potential in the thesaurus, for us a suitable level.

Conversion of our thesaurus was accomplished quite readily. Two short programs were written. The first, for one-time use, stripped the old codes from the thesaurus tape, apportioned the letters, and calculated and entered the check digit. The second, for repeat use, is a screen by which all new punch card input is reviewed and any card with an error is rejected and listed. The error lists are checked by hand, the erroneous cards are located, corrected, and rekeypunched. All proofed and corrected cards are then accumulated for a master file update. (A simple presorting is of considerable help later in locating erroneous cards).

Advantages obtained from the use of check digits were: Errors in term codes could be found and corrected easily, thus only "clean" input need be used for master file updating. Verifying—i.e., duplicate punching—could be eliminated. Only those cards in error, in fact, need be repunched. (I might add that we had found substantial abuse of the verifying punch key.) The increased confidence in catching mistakes led to improve keypunching speed. Thirdly, we rediscovered that term codes of one letter and six digits were easier to remember in copying our old seven-digit code, Bell Telephone's position notwithstanding.

A second and more novel approach to eliminate transcription error was the creation of a "rubber stamp thesaurus". The stamps also speeded term code assignment to the tracing sheets.

Figures 4 and 5 illustrate a typical term stamp, and the rack in which they are retained with use by the code clerk. Each lettered check digit term is converted to a conventional rubber stamp strip ($\frac{3}{16} \times \frac{15}{16}$ inch) which is mounted on the left end of a small pine block ($2\frac{13}{16} \times 1$ inch). To the front edge is affixed the label in English; to the back edge is glued a narrow strip of 18 gage galvanized iron. The stamps are retained in a unique compact rack comprising magnetic panels vertically arranged in four columns per hinged leaf, with 10 leaves per rack for a total of more than 3000 stamps. Each magnetic panel is about three-quarters as wide as the length of the stamp and is mounted so that the left edge is raised from the plane of the leaf. This raised edge serves as a fulcrum so that the clerk with a simple deft push of a pencil or a long fingernail can dislodge the rear end of the stamp directly into the grip of the remaining fingers of the hand, a position just right for

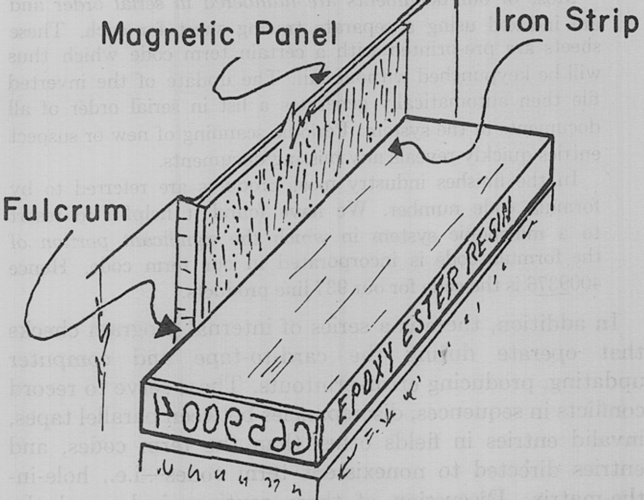


Figure 4. Typical stamp and magnetic mount

Figure 5. Stamp rack in use

stamping the term code on the document or tracing sheet.

We found from experience that a certain number of the term codes were used quite frequently in indexing and hence justified the expense of stamp preparation. Repetitive "look-up, write-down" is avoided and transcription is both speeded up and errors minimized.

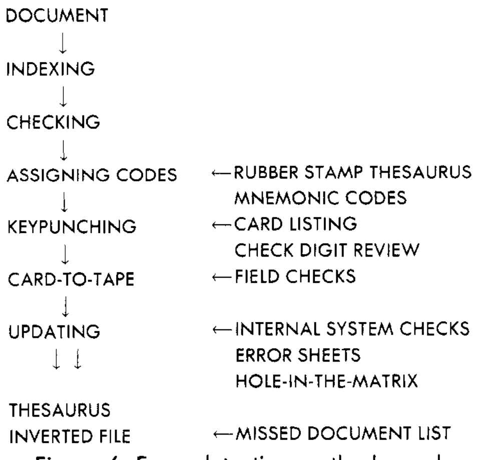
A third technique which we use in combination with the others is termed "hole-in-the-matrix." This phrase is a way of recognizing the fact that only 1 out of every 200 spaces in the thesaurus actually is dedicated to term use—i.e. 13,000 out of 2.5 million. Should a term code error be undetected by the other screens, it still has only one chance in 200 of coinciding with a term already established and thus escaping notice.

The above routines serve only the term codes. Other routines were required for erroenous or missing document numbers and related independent information:

> Obvious, of course, is simple 80/80 listing of the card decks as they are punched. Scanning of the lists readily identifies trash cards, displaced fields, and unusual document numbers. In addition, despite use of automatic drum cards in punching, a certain percentage of cards have to be free-punched anyhow, with attendant field discrepancies.
>
> Most of our documents are numbered in serial order and are indexed using a separate tracing sheet for each. These sheets are pre-printed with a certain term code which thus will be keypunched without fail. The update of the inverted file then automatically generates a list in serial order of all documents in the system. Periodic scanning of new or suspect entries quickly reveals any missing documents.
>
> In the finishes industry most products are referred to by formula code number. We have found it helpful to resort to a mnemonic system in which the significant portion of the formula code is incorporated in our term code. Hence 4009376 is the code for our 937 line products.

In addition, there is a series of internal program checks that operate during the card-to-tape and computer updating, producing error-printouts. These serve to record conflicts in sequences, discrepancies between parallel tapes, invalid entries in fields other than the term codes, and entries directed to nonexistent term codes—i.e., hole-in-the-matrix. Discussion of these routines is beyond the immediate purpose and scope of this paper and would seem to be largely specific to the design details of our particular system.

In recapping, the techniques used in the order of their use along the flow diagram in Figure 6 are:



DOCUMENT
INDEXING
CHECKING
ASSIGNING CODES ←RUBBER STAMP THESAURUS
MNEMONIC CODES
KEYPUNCHING ←CARD LISTING
CHECK DIGIT REVIEW
CARD-TO-TAPE ←FIELD CHECKS
UPDATING ←INTERNAL SYSTEM CHECKS
ERROR SHEETS
HOLE-IN-THE-MATRIX
THESAURUS
INVERTED FILE ←MISSED DOCUMENT LIST

Figure 6. Error detection methods used

EVALUATION

As noted, before adopting this series of techniques, up to 10% of our input was suspect and suffered review. This required in the order of six man-weeks to correct after an updating. Furthermore, the uncertainty of entries in the inverted file was a source of concern in searching. Let me summarize by saying that at present after 2½ years of experience, I am unaware of any significant errors in the system and I do not anticipate that there are any. The following data on errors corrected in our last two updatings are of interest (Figure 7).

	April	August
Cards punched (30 punches/card)	7550	4600
Errors at listings	[a]	[a]
Errors at check digit	244	257
Errors at card-to-tape	10	2
Errors at update	[b]	[b]

[a] Records of errors corrected after listing have not been kept and are the individual responsibility of the keypuncher to find and correct. [b] Errors at update because of their complex nature are only partially pertinent here. However the total number of pages of error printout at this stage has been reduced about to one fifth of that encountered before the check digit and stamp programs. Virtually all of the errors encountered now are either "intellectual—i.e., bad guesses on relationships or terminology, or from critical cards in a sequence not being processed.

Figure 7. Error experience

Using the combined techniques reported here the time to review and to correct an update has been reduced to about one man-week and the reliability of searches is improved to about as good as the primary indexing. This performance has been obtained despite the fact that we have had major personnel turnover, hence training problems, in the group in the 4 years of operation. The group has included three different full-time keypunchers (and many part-time borrowed ones) and five different technical indexers.

What are the chances of errors getting through these screens? This is somewhat like asking the guide in the Carlsbad caverns: "How much of the cave has yet to be explored?" I have made no statistical calculations nor rigorous analysis, but on a practical basis I do not know yet of any significant error. The check digit system of course is not infallible and an occasional error can escape the routine. Figure 8 illustrates four situations where the check digit is ineffective.

	Found	Should Be
0,9 Transposition	G650904 ↑	G659004 ↑
One-three transposition	G956004 ↑	G659004 ↑
Displaced fields	G| 6590046| 0 _| ↑ ↑	| G659004| 60|
Letter value errors	C604004 ↑ ↑	G640004
	V410000 ↑ ↑	Y140000
	V300003 ↑ ↑	Y030003
	F076002 ↑ ↑	E067002

Figure 8. Check digit failures

I must emphasize, however, that occurrence of these has been rare and in each instance the error was found nonetheless because of subsequent "hole-in-the-matrix" situations.

The weaknesses of the check digit system shown seem to fall into two categories: (a) single error transpositions that are allied with peculiarities of the number theory or the design of the modulating routine—e.g., the multiplier, and (b) random paired compensating errors. In the first category, as shown by the arrow, no matter which digit of the pair 0,9 is multiplied by 2, the net contribution to the check digit is still the same—i.e., 9. (See routine on Figure 2.) A short calculation will show that this is true for any single digit multiplier for one pair of digits of a number system to any base. In the same sense, the net from either form of a one-three transposition is the same. Fortunately this type of error rarely occurs.

The second category, random paired compensating errors, do occasionally occur. In the displaced field example above, the loss of the value of "G" by its leftward displacement out of the field is exactly compensated by the invasion of a 6 from the field to the right. Such displacement errors are usually detected first at the card-to-tape routine that locates blank columns. However, in those codes where three or four zeros occur in a string, the keypuncher may also unwittingly be enticed into putting in an extra zero, thus negating the blank column check.

Letter value errors due to careless writing or faulty recognition are not uncommon: C for G, V for Y, Y for X, U for V, and E for F have been noted. However, to escape detection a second specific compensating error must also have been made, and in practice this rarely seems to occur. Figure 9 shows pairs of numbers which are the respective net check digit contribution that would be calculated in a simple x,y/to y,x transposition. Wherever the difference between members of a pair equals the change in value due to the letter being substituted for another (that is, in a final check digit calculation of a code) then the errors obviously escape detection by the described system. In the example, the difference in net contribution between "C" and "G" is 13 − 17 = − 4. The net contribution of the substitution of 4,0 by 0,4 is 6 − 2 = + 4.

Let me emphasize, the above errors occur infrequently, and in our scantily filled thesaurus—e.g., 13,000 terms out of a total of 2.5 million, those that do escape check digit routines stand a high chance of being found by the "hole-in-the-matrix."

Cost of this error control system, I believe, is insignificant compared to the benefits derived. Keypunching is done only once; only erroneous cards are repunched. Both the check digit and card-to-tape are inexpensive: roughly$1/per thousand cards. At $10 to $15 per update, this is substantially less than the alternative cost of screening and printing-out errors during the actual run at $200/hr. computer time. Cost of the internal system checks seems to be about $100 per update. When setting up a new system, the additional effort, of making the thesaurus one based on check digits, is small—say $50 to $100.

	X Digit								
Y Digit	0	1	2	3	4	5	6	7	8
1	8,9				First			Second	
2	6,8	5,6			Check			Check	
3	4,7	3,5	2,3		xy			yx	
4	2,6	1,4	0,2	9,0	$\underline{2}$			$\underline{2}$	
5	9,5	8,3	7,1	6,9	5,7				
6	7,4	6,2	5,0	4,8	3,6	2,3			
7	5,3	4,1	3,9	2,7	1,5	0,2	9,0		
8	3,2	2,0	1,8	0,6	9,4	8,1	7,9	6,7	
9	1,1	0,9	9,7	8,5	7,3	6,0	5,8	4,6	3,4

Figure 9. x,y Transpositions, effect on digit check

ALTERNATE SYSTEMS

No extensive comparison with alternate systems has been undertaken. However, reactions by retrieval people to the described system has been interesting. One of the prize comments was, "We don't make mistakes, hence don't need such a system." I can make no rebuttal to such claims of perfection. I suspect such claims might better be used for lofting thermal balloons.

Another reaction has been, "We index and update using the English words directly, thus avoiding a code system entirely." An English word system of course has its merits, especially if chemical terms are handled by a separate non-word-code system. However, I believe there is as much or more effort in looking up in the thesaurus to verify the correct spelling, or the accepted suffix for a given term, as there is in setting up or using the seven-digit code for the same term. In addition, there is no assurance that the final English input to the update will retain the required accuracy without meticulous visual proofing. Further, any word involving more than seven characters is at an increasing disadvantage both in transcription and keypunching, and in the complexities and poor utilization associated with programming of variable-length word fields. In smaller systems where the thesaurus includes both general as well as chemical and related multi-stem technical terms, the seven-character term code is much to be preferred because of the flexibility in choice of English. Indexing, transcription, keypunching, and thesaurus management is accordingly simplified. Minor spelling errors, for example in a 35-character English term that could easily go undetected, and lead to misplaced entries and all the restoration effort this implies, are completely avoided in a check digit system. In effect, the check digit system can correctly handle sloppy work with minimal effort. Indexing input is easy to correct and input does not get in the system unless it is correct. An update does not have to be made with anything but high quality input.

In summary, the advantages of a check digit error control procedure are:

A thesaurus of combined general and technical terms is practical.
Sloppy work from training and personnel turnover problems can be readily corrected.
Eliminates duplicate punching (verification).
Speeds initial keypunching.
Error correction at card-to-tape is cheaper than at updating.
Post update clean-up is quicker and cheaper.
Searches are more likely to be thorough.