

RALF—A New Software Package for the Whole Complex of Punched Card Oriented Documentation*

K. LOCH and W. NUEBLING**
E. Merck, Darmstadt, Germany

Received July 1, 1973

A new, universally applicable computer program is described for the whole complex of punched card oriented literature documentation. Its characteristics are: modular construction, therefore easily adapted to user-specific problems; extremely high searching capacity; comprehensive and elastic syntax makes possible new unconventional methods of evaluation.

It was in 1964 that we at Merck turned for the first time to the computer as a means of making use of the documents accumulated by the Pharma Documentation Ring.

From 1958 until then, 700,000 punched cards had been accumulated for the processing of which, using an IBM 1401, we had written our own program. After the changeover to the IBM 360, we more or less copied this program, but we did not match it to the considerably increased potential offered in principle by the new system. The desire to make use of this increased potential, even more so following the changeover to System 370, the wish to shorten working times and to make best use of all the variants given by the combination of our documentation methods and computer techniques led us to develop a completely new type of software package, named RALF (Rapid Access to Literature via Fragmentation Codes).

Data held at present amounts to about 2.7 million tape records, including material from the Pharma Documentation Ring and from Ringdoc, Vetdoc, Pestdoc, Farmdoc, Agdoc, CPI, our own reaction index, the US steroid index, together with a wealth of Merck-internal material.

PLAN

Often in serial systems considerable CPU times are encountered with high costs.

Our aim was to develop a complete program system for data maintenance and searches which should:

- (a) Incorporate all the presently known possibilities offered by serial systems
- (b) Be able to answer completely new types of questions
- (c) Be at the same time extremely efficient
- (d) Be inexpensive
- (e) Be easy on the user

PROGRAMMING

Capacity. All programs are written in the assembler language of the 360 system. The program system is built up in a modular construction, meaning that alterations may be performed simply.

The program is designed for OS, but can be adapted also for DOS.

All program parts can make do with a maximum of three tape units, one 2314 disc unit, and an 80K core stor-

age unit; the capacity can be improved still further by the use of a larger store.

Our program goes the way of a compilation of the inquiry text input into an optimized machine program. The program consists of 15 phases, each overlapping the preceding one; only one common range with control data remains as an entity.

The program flow is data-dependent; only those phases and modules are loaded which are actually required.

The following capacities are achieved:

Data on hand 2.7 million tape records

Turn-around time 40 minutes

Number of inquiries 25 (increase easily possible)

CPU time about 3 minutes + $\frac{1}{2}$ - 1 minute per inquiry, depending on complexity and amount of output

The total turn-around time occurs only during a search of the *whole* of the stored data—i.e., without skipping subfiles.

The question "inverted file or serial" has all but become a question of weltanschauung. Both data index structures have their disadvantages as well as their advantages.

Inverted File

- Advantages:** high search speed
possibility of dialog
possibility of a step-by-step search
- Disadvantages:** expensive periphery
data is no longer arranged in the original form
complicated and expensive updating

Serial

- Advantages:** easy management of data index
data in original form
inexpensive periphery
- Disadvantages:** turn-around time and, nearly always, CPU time extremely long

We believe that we have combined the advantages of both types; only the dialog cannot be implemented.

DATA INDEX FORMAT

The data index format has been so designed that all punched card oriented original data may be processed—i.e., including firm-internal data, no matter what the structure is (Figure 1).

The tape record consists of 5 main parts

- (a) The complete card input in condensed form
- (b) A card index marker
- (c) A year marker
- (d) A program-dependent control field
- (e) The so-called identification characteristic (ID)

*Presented before the RINGDOC subscriber meeting, Derwent Publications Ltd., London, June 6/7, 1973.

**To whom questions should be addressed.

Length byte

120

1 1

4

12

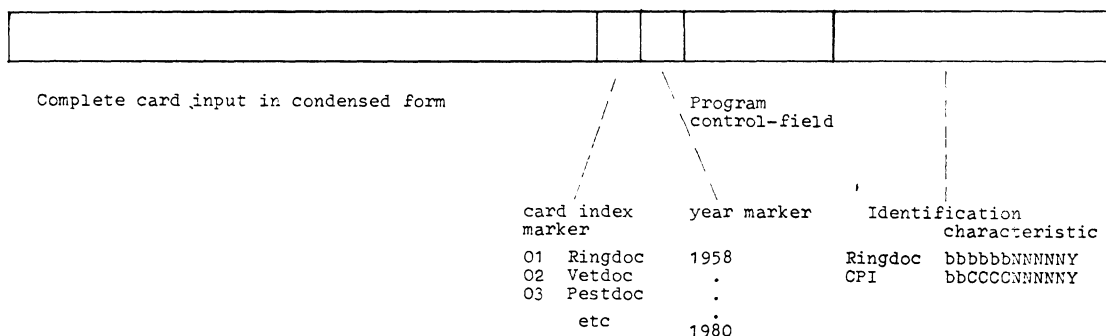


Figure 1. Data format

INQUIRY	0004606
DEPARTMENT	CHEM LIT-TEST
INQUIRER	WOERTH
CLIENT	DR MUELLER REA
DATE OF INQUIRY	23.05.73
INTERNAL INQUIRY NUMBER	57/67
CARD FILE	CPI * RING
CHECKED	BO
CODE	NOT SPECIFIED
SUBJECT	3-PYRIDYLFLUORON
TYPE OF SEARCH	CODE-SEARCH
	NO LINKING OF QUESTIONS
TYPE OF OUTPUT	NORMAL
OUTPUT LIMIT	200
DIAGNOSIS FORMAL	NO ERRORS
CODING	NO ERRORS

DEPARTMENT	INQUIRY 0004606	PAGE 01
		PAGE 02
CHEM LIT		
SEARCH IN DOCUMENTATION CARD FILES	RUN NO.	
511	DATE 25.05.73	PAGE 542

Figure 2. Cover sheet 1

STATISTICS	
TOTAL	2 348 726
RUN-THROUGH TAPE RECORDS	2 049 623
REFERENCES FOUND	20
MODULE LENGTHS IN BYTES	576
RUN-THROUGH VOLUMES + CARD FILES	
K01,K04-21,K32,K35,K37,K38	
CODING	

K01-21*02&*020*021*031
 *05&*06&*071*075*080
 *094*10-*102*104*130
 *18-
 *%032*040*046*063*101*182*25-
 *251*257*062,033*064*09&*13&
 *18&*180□*#%01&.017.02-022.023□
 .K32-38*314*176*59-*61&*62-*62
 0*493*#%01-.010.011.012.015
 .016.017.018.019□

INQUIRY 0004606	PAGE 02
DEPARTMENT CHEM LIT	PAGE 03
SEARCH IN DOCUMENTATION CARD FILES	RUN NO.
511	DATE 25.05.73
	PAGE 544

Figure 3. Cover sheet 2

This construction has the following advantages:

1. Source data are retained in their original form, update errors, etc., are completely uncritical; questions may be put concerning the whole card
2. The type of index and year are given explicitly; they can be queried directly
3. The identification characteristic is extracted from the card contents. In a few cases (e.g., Ringdoc) only the serial number and year marker are given, though any other information may also be included (e.g., company code, or firm-internal information)
4. The source data are compressed so that there is no redundant information; one byte contains the information from 8 punched card positions, an important prerequisite for an effective optimization of searches.

STRUCTURE OF THE DATA INDEX

To obviate the need to search through all the material, we have divided the whole data stored into about 20 subfiles. On each subfile there is a logically connected intermediate field of the total information held.

Both the data maintenance programs and also the search program automatically calculate

(a) By means of data to be added or altered during the course of updating

(b) By means of the card indexes and years addressed in the individual inquiries

which subfiles should be completed or searched; the remaining subfiles are automatically skipped. Using this technique, even an amount of data as large as that we deal with can be easily handled.

At the same time we achieve the following.

Every addition, correction, or deletion in the data held affects only a small part of the total volume, being carried on a physically separated data carrier. Movements within the data are thus limited to a fraction of the total data.

We have arranged our data according to topicality, that is to say, the most recent abstracts are at the beginning. This arrangement is normally critical with large amounts of data, since even where there are a few new additions the whole vast remainder must be pushed to the rear.

Apart from this "let out," the search program also has the following possibilities: Only those inquiries are processed in the individual subfiles which are actually relevant to a particular subfile; we call this possibility "preselection." This contributes to a very high degree to the speed of the program.

STATISTICS

```

RUN-THROUGH TAPE RECORDS  INQUIRY PART1 191 623
REFERENCES FOUND           INQUIRY PART 2 191 617
REFERENCES FOUND           INQUIRY PART 1  1878
                           INQUIRY PART 2   771
                           TOTAL             21
MODULE LENGTHS IN BYTES   INQUIRY PART 1   170
                           INQUIRY PART 2   358

```

INQUIRY PART 1

RUN-THROUGH VOLUMES +CARD FILES

K01,K04-21
CODING

538

INQUIRY PART 2

RUN-THROUGH VOLUMES + CARD FILES

K01,K04-21
CODING

672*686*690,513,010*02&*06&
*060*075*10.*103*104*105*106*107
*115*125*13&*130*18&*181*403
*#01&.017.02-.020.021,022,023

INQUIRY 0003596 PAGE 02

DEPARTMENT MED LIT PAGE 03
SEARCH IN DOCUMENTATION CARD FILES RUN NO..
.00 DATE 27.10.72 PAGE 272

Figure 4. Double-linking

SEARCH POSSIBILITIES

We can search in two sections of the tape record

- (a) in the ID characteristic
- (b) in the subject matter section

Access is gained to the year and card index data in both cases; the control field is only of importance for internal programming purposes.

Search in the ID Characteristic. It is possible to search both for individual abstracts and for whole years or intermediate fields, likewise for individual pieces of information, such as company code, or parts thereof, or all abstracts with the final number 13, etc.

Search in the Binary Section. Since the original is present in its entirety in the tape record, the whole card can be inquired of; simultaneously, access is gained to the respective year and index characteristic.

Possible Search Strategies. The normal case would be that of a simple search. Figures 2 and 3 show the output.

Whenever an abstract contains various sets of subject matter which cannot be coded on one card, a separate card exists for each set of subject matter. It is not possible to link these components in a search with "AND," since although one of the components turns out to be positive, all the others do not. Linking the components with "OR" means that each abstract, which is only one of these components addresses, is recognized as positive. In such cases, what is searched for are abstracts in which both components A and B occur. The output from one of these double-linked inquiries can be seen in Figure 4.

The additional time required per inquiry amounts to about 5%. The printout gives the number of abstracts found, the amounts of A and B, and the average amount. Also, triple-linking of subject matter is possible. In this case the additional time required as compared with the normal feed-in amounts to about 8%.

If the same abstract is present in two different codes (e.g., Ringcode and CPI), then it is possible to search in both indexes with only one inquiry, with part of the abstracts overlapping. There is no point in printing out

PRINTOUT OF RESULTS OF INQUIRY 0004606

```

CPI-B/CHEMISTRY 1972 STAD D 62156 T
                  /SPI D 59702 T
                  UPJO D 22547 T
                  MERI D 21839 T
CPI-C/CHEMISTRY 1972 STAD D 62156 T
CPI-E 1/E 3      1972 MONS D 22322 T
CPI-B/CHEMISTRY 1971 ALZ* D 70343 S
                  DOWO D 63065 S
CPI-E 2          1971 DUPO D 71860 S
CPI-E 1/E 3      1971 DOWO D 63065 S
                  DUPO D 39009 S
CPI-E 2          1971 BELH D 27196 S
CPI-E 1/E 3      1971 EAST D 21867 S
CPI-B/CHEMISTRY 1970 /NOU D 94110 R
                  BOOT D 23193 R
                  UMIS D 07590 R
CPI-E 2          1970 CIBA D 80978 R
                  GEVA D 62275 R
                  DUPO D 24590 R
RINGDOC          1966      12484 F

```

INQUIRY 0004606 PAGE 03

DEPARTMENT CHEM LIT PAGE 04
SEARCH IN DOCUMENTATION CARD FILES RUN NO..
511 DATE 25.05.73 PAGE 544

Figure 5. Normal printout

twice those abstracts which are present in both indexes. The program automatically eliminates those abstracts in one index which are also present in another code. In every case, the index with the lower index number is printed out.

Output. The output is sorted within the individual inquiries as follows:

1. Years (the most topical ones first)
2. Card indexes (progressing according to index numbers)
3. According to serial numbers; the most topical (highest) are printed out first

Structure of the computer printout:

It is possible to control separately both the quantity and the nature of the printout for each inquiry.

1. Quantity. The program in each case prints out only the 200 most topical abstracts (standard limit). If required, this limit can be lowered or raised by the inquirer.
2. The type of output can, all according to the particular task, take the most diverse forms, again separate for each inquiry.

Presently we can choose between 12 types of output, of which we mention the following:

- (a) Output of abstract numbers (Figure 5)
- (b) Complementing with bibliography from codeless scanning
- (c) Card printout (Figure 6)
- (d) Position printout
- (e) Card punch

Special Functions.

1. Statistics concerning the material found. The documentation search program further permits a statistical survey of the contents of the binary part of relevant abstracts; this is done in the form of a bit count. By this means it is possible to count both individual columns and also column intervals. By examining the bit count, it is possible to form conclusions regarding the quality of the code used. Actual contents and an even distribution of the positions are in general a direct measure of the selectivity of a code. This criterion of even distribution is computed by the program and printed out in column or position extremes. Simultaneously, the percentage share of a particular punch in the material found relevant for the original inquiry is also held fast. This permits a statistical analysis

PESTDOC

1971

83480 L

```

          9 C      F -Z          6 9E 8 382
-----
12 1 .XX.....X.....X.....X.....X.....
11 1 .X.....X.....X.....X.....X.....
0 1 .X.....X.....X.....X.....X.....
1 1 .X.....X.....X.....X.....X.....
2 1 .X.....X.....X.....X.....X.....
3 1 .X.....X.....X.....X.....X.....
4 1 .X.....X.....X.....X.....X.....
5 1 .X.....X.....X.....X.....X.....
6 1 .X.....X.....X.....X.....X.....
7 1 .X.....X.....X.....X.....X.....
8 1 .X.....X.....X.....X.....X.....
9 1 .X.....X.....X.....X.....X.....
-----
000000000111111112222222223333333334
1234567890123456789012345678901234567890

```

```

          97      OE          83480J
-----
12 1 .....X.....X.....X.....X.....X.....
11 1 .....X.....X.....X.....X.....X.....
0 1 .....X.....X.....X.....X.....X.....
1 1 .....X.....X.....X.....X.....X.....
2 1 .....X.....X.....X.....X.....X.....
3 1 .....X.....X.....X.....X.....X.....
4 1 .....X.....X.....X.....X.....X.....
5 1 .....X.....X.....X.....X.....X.....
6 1 .....X.....X.....X.....X.....X.....
7 1 .....X.....X.....X.....X.....X.....
8 1 .....X.....X.....X.....X.....X.....
9 1 .....X.....X.....X.....X.....X.....
-----
444444444555555555566666666667777777778
1234567890123456789012345678901234567890

```

Figure 6. Card printout

of structure—effect relationships, insofar as is possible within the limits set by the code. This result can also be represented graphically. Figure 7 shows one page of the statistical table, and Figure 8 shows the graphical representation. All special functions of this type are incorporated into the printout for the appropriate inquiry.

With the aid of a special program, it is possible to compare punch pictures. The reply to the question "Which single compounds were described in 1971 as cytostatics?" is simplified by the fact that each punch picture occurring is registered per program only once on its first appearance. If the same punch picture occurs for a second or third time during the course of a search, then the program makes allowance for it to be skipped. The final result is a collection of abstract numbers of which each individual one refers only to a quite definite molecule. In conjunction with the printout of the punch picture, it is relatively easy to establish which compounds have been addressed by the computer as "new individuals." Together with the bit count, this program allows an interesting insight into effect displacements with structural alterations. Of course, coding errors here have a particularly disturbing influence (Output Figure 9).

2. Step-by-Step Search. There are two possibilities for the performance of so-called step-by-step searches.

(a) By tracing the logical decision pattern together with marking of part components

(b) Independent expansion or contraction of the inquiry with the aim of reducing the output to a previously defined limit.

We have given preference to solution (a).

Our search program constructs a decision pattern from the inquiry text; this pattern will proceed by the shortest possible route. Within this decision pattern, it is possible to set counters at interesting positions to record through-runs together with results.

The counter number and number of through-runs are

Result of Bit Count before Structure Comparison

COL-UMN	POSITION	COUNT	% of COLS.	% of TOTAL	% of TAPE RECORDS	TOTAL
1	12	519	3.690	0.212	5.996	
	11	114	0.810	0.046	1.317	
	0	2264	16.097	0.928	26.158	
	1	832	5.915	0.341	9.612	
	2	4	0.028	0.001	0.046	
	3	525	3.732	0.215	6.065	
	4	2	0.014	0.000	0.023	
	5		0.000	0.000	0.000	
	5	7	0.049	0.002	0.080	
	7	1142	8.120	0.468	13.194	
	8		0.000	0.000	0.000	
	9	8655	61.540	3.550	100.000	14064 5.770 %
2	12	5115	37.874	2.098	2.797	
	11	2574	19.059	1.056	29.740	
	0	2723	20.162	1.117	31.461	
	1	1825	13.513	0.748	21.086	
	2	382	2.828	0.156	4.413	
	3	148	1.095	0.060	1.709	
	4	161	1.192	0.066	1.860	
	5	259	1.917	0.106	2.992	
	6	254	1.880	0.104	2.934	
	7	64	0.473	0.026	0.739	
	8		0.000	0.000	0.000	13505 5.540 %

INQUIRY 0004400 PAGE 04

DEPARTMENT CHEM LIT

SEARCH IN DOCUMENTATION CARD FILES RUN NO. 511 DATE 25.05.73

PAGE 05

PAGE 07

Figure 7. Bit counting

RALF—A NEW SOFTWARE PACKAGE FOR CARD ORIENTED DOCUMENTATION

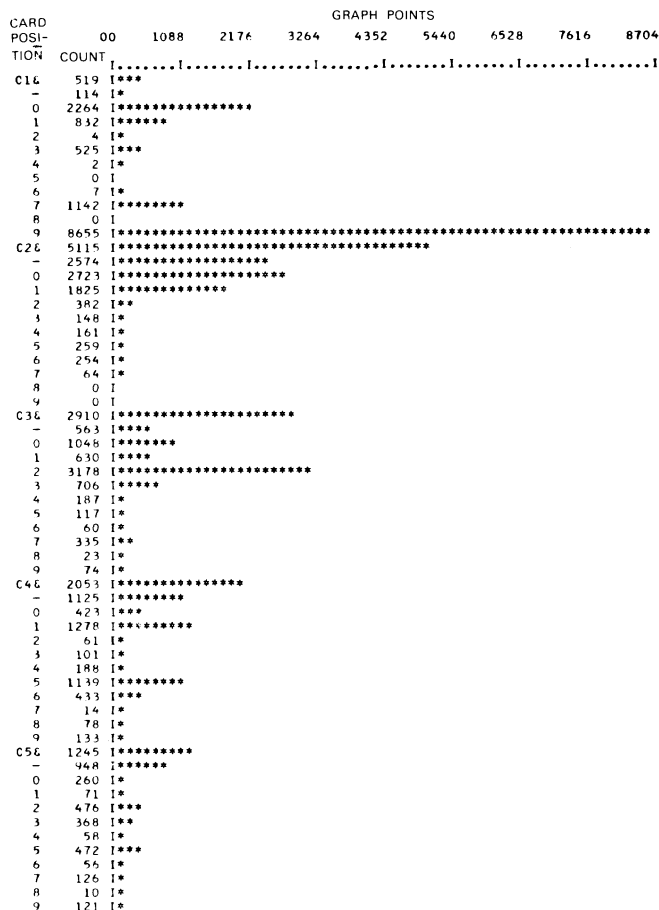


Figure 8. Graphical printout

given in the inquiry printout, thus giving rise to a continuous picture of how the searched material reacts towards the inquiries put.

Simultaneously, it is noted on the abstract number, once found, which counter was run through last (see Figure 10).

The simple inquiry at the beginning of the sheet generates the decision pattern which follows. Counters were placed at the intersections.

There are two unpleasant cases with which the documentation specialist can be confronted:

1. No output
2. Too much output

In both cases it is possible to establish immediately, by means of the counter record, which sections of the inquiry have imposed unduly harsh limitations or which have contributed nothing to the solution of the problem. It is thus possible to make targeted alterations to the inquiry, instead of feeling one's way through.

This type of usage will also be particularly advantageous in error searches and for the training of new personnel.

The marking of the abstract numbers fed out with the adjacent counter reading can be used to advantage to associate the abstract with certain sections of the inquiry, for example, in a search for a chemical structure with three medical side effects.

If the side effects are marked with a counter, the following is obtained:

1. A record of what side effects are associated with this structure, and how often
2. In the abstract number given, an indication as to what side effects the abstract describes.

Syntax. The syntax of the search program, as compared with that of the already well-known programs, has been considerably expanded.

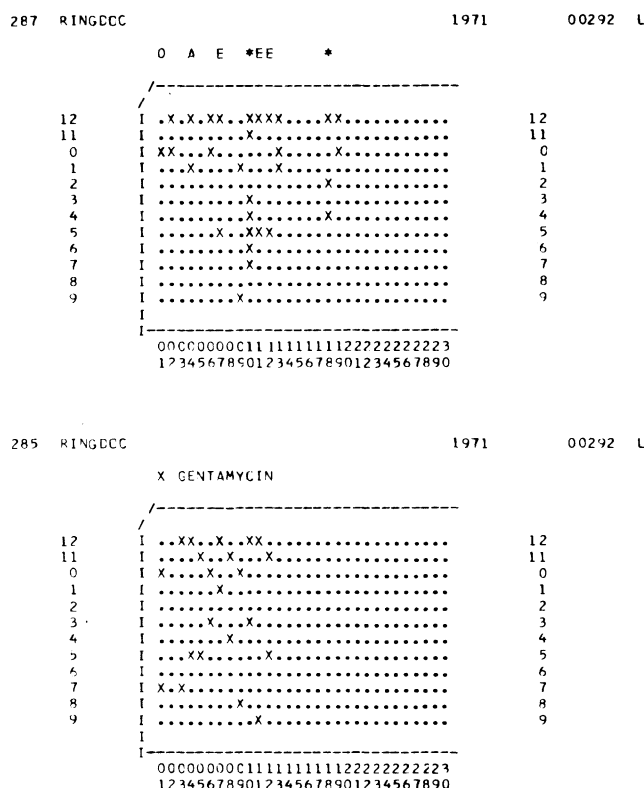
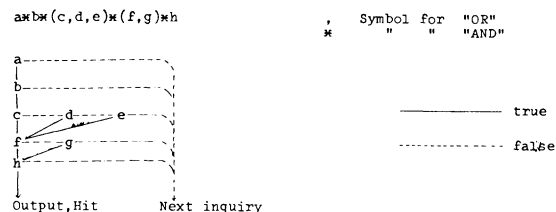
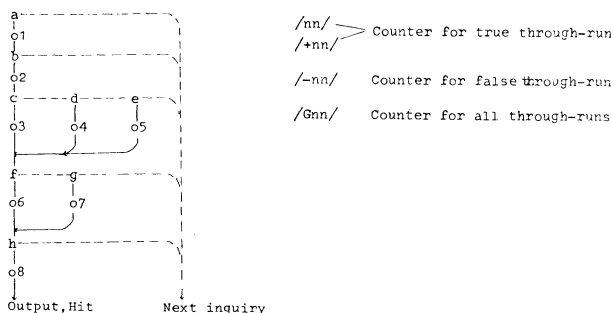


Figure 9. Structure printout



Counters of step-by-step search included

a/o1/b/o2/c/o3/d/o4/e/o5/f/o6/g/o7/h/o8/



Counter number	through-runs	% of records
o1	26.024	16.67
o2	5.262	.
o3	4.03	.
o4	8.943	.
o5	12	.
o6	1.468	.
o7	67	.
o8	36	.

Figure 10. Step-by-step search

[illegible]

Figure 11. Inquiry cover sheet

The signs for the connectors “And” and “OR” and negation are elementary, likewise opening and closing brackets.

The brackets and the negation are not subject to any restrictions, either in their absolute number or in their progression.

1. Card and year outputs

It is possible to refer back to the card index marker and year marker in the search printout, to search, for instance, for:

individual years according to indexes	Knn or Jnn
left open intervals, e.g.,	KBnn or JBnn
right open intervals, e.g.,	KAnn or JAnn
closed intervals, e.g.,	Knn-nn or Jnn-nn

where K = subfile, J = year, B = ending at, A = beginning at.

2. Putting questions concerning punched card positions
 - (a) In trivial terms, the normal data reads SSP, where SS gives the card column and P gives the position.
 - (b) Whole card columns can be dealt with.

Examples:

The expression “occupied column” addresses all 12 positions. The stipulation is that at least one position must be occupied or (in the case of negation) no position should be occupied.

The expression "total column" stipulates that all the positions in one card column must be occupied.

These data are of particular advantage in conjunction with negation.

(c) Illness numbers can be searched for by simply placing the latter M in front, e.g.,

M28901 (= hyperlipoproteinaemia),
that is to say, without giving additionally the columns
and positions.

(d) Position intervals can be inquired of simply by giving the upper and lower limits, in conjunction with the required connector; example: Any of the following positions should be occupied: 350 – 359 (.)

(e) The specification "cleartext" permits trouble-free coding of even fairly long cleartext passages. Only the initial column of the text and the cleartext, set in inverted commas, is given, e.g., 03'PHOSPHOMONOESTER'

In searches for inorganic substances—e.g., quite generally for inorganic zinc compounds—the search is often quite involved, since the series of symbols can lie in various card columns. It is possible to have the cleartext search commence automatically in various columns, or “iterate”

For example:

We are searching for ZN in columns 3-10. You iterate simply I03-10'ZN'.

Input. Input of the inquiry ensues by punched cards; the format and the way of exploiting the punched card can easily be modified on account of the modular construction of the program.

Figure 11 shows the cover sheet of the inquiry form.

PROSPECT

Individual programs such as
 plausibility testing of new additions
 testing for completeness
 simulation of the step-by-step search
 bibliography
 searching via Codeless Scanning

are being tested at the moment and will be implemented in the near future.

The currently available program parts—namely,

1. searching with the special functions described here
2. updating together with statistics program

have been operating for some considerable time now and, quite apart from the saving in costs, they have afforded us a great new potential for the qualitative improvement of our work.