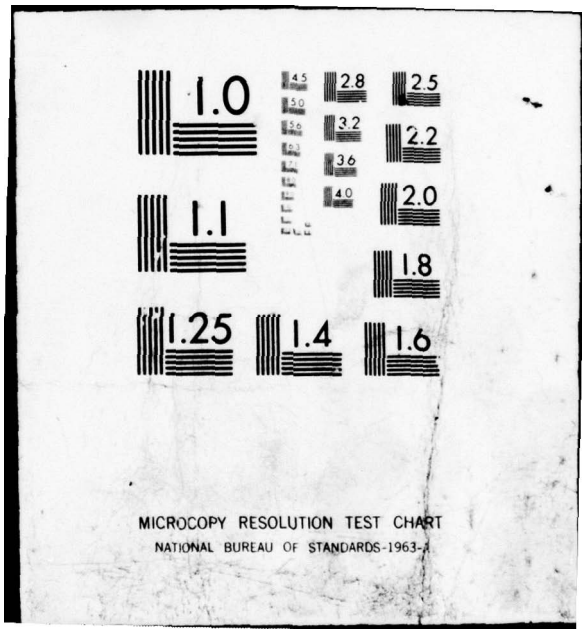
AD-A056 641 AIR FORCE FLIGHT DYNAMICS LAB WRIGHT-PATTERSON AFB OHIO F/G 5/2
INFORMATION STORAGE AND RETRIEVAL (A BRIEF INTRODUCTION),(U)
SEP 74 H B THOMPSON
UNCLASSIFIED AFFDL/TST-P-74-1 NL

| OF |
AD
A056641

END
DATE
FILMED
8-78
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFFDL/TST-P-74-1 LEVEL II

AD A056641

AD No. ____
DDC FILE COPY

INFORMATION STORAGE & RETRIEVAL

(A BRIEF INTRODUCTION)

SEPTEMBER 1974

CL

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER (14) AFFDL/TST-P-74-1 ✓	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) (16) Information Storage and Retrieval (A Brief Introduction),		5. TYPE OF REPORT & PERIOD COVERED
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) (10) H. B. Thompson	(11) Sep 74	8. CONTRACT OR GRANT NUMBER(s) (12) 24p.
9. PERFORMING ORGANIZATION NAME AND ADDRESS AFFDL/TST Wright-Patterson Air Force Base, Ohio - 45433		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS AFFDL/TST Wright-Patterson Air Force Base, Ohio - 45433		12. REPORT DATE Sept. '74 13. NUMBER OF PAGES 20
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) U 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Information processing, coordinate indexing, storage, retrieval, search strategy, KWIC, key-word, light-coincidence, edge-notched cards, logic, posting index, Miracode, Computer retrieval, thesaurus, sources, information services, terminals

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

During this century classical library cataloguing has become inadequate for the flood of technical literature being produced. Coordinate indexing can include all subjects dealt with and can be stored in various ways. The advantages and disadvantages of each technique for input, storage, and retrieval are given. Storage and use of paper copy vs microform and different forms of output are discussed. Special sources for special or less frequently needed material are given.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

AFLC-WPAFB-OCT 74 500

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

78 07 13 039

012070

CL

TABLE OF CONTENTS

LIST OF FIGURES

Information Storage and Retrieval

Pre WW II Information

In school and for our own interests out of school we have all used libraries and generally found them very helpful. Information on almost any not too specific subject is available and usually conveniently. This has been the case for decades and perhaps centuries. It seemed adequate in the first part of this century.

There were other sources of both information and inspiration - journals, abstract journals, symposia and conferences, and knowledgeable people. These are all still available and much used. Because we are familiar with them they are often very satisfying.

Technology Explosion

In the 40's a new development emerged. Scientific research multiplied, the results were recorded, and these records became unmanageable by established means. It became impossible to do any significant research and keep up on what others were doing in the general area. Accepted methods didn't subdivide the literature adequately. In many cases the combination of fields previously not closely associated promised desirable results. Since the relation had not been observed before, cataloguing was not done with this retrieval in mind. How were these needs to be met?

Coordinate Indexing of Documents

The concept of coordinate indexing was advanced and has grown since the late 40's. Fundamental to this indexing was the thought that all subjects on which information was given should be indexed. High specificity could be provided, retrieval could be on any combination included, and most unrelated material should be avoided. The idea is superb, theoretically achieveable, and has been assumed by many to be a reality. However, all documents of potential value to an organization must be in their file if they are to be retrieved and they must be well indexed. Having or being able to obtain virtually all pertinent documents is possible but getting a desired item may be time consuming and in some cases expensive.

Adequate indexing is another question. Many have believed it is only necessary to say, "Index what's important." What is important to one individual is not important to everyone. Even major automated coordinate retrieval systems, like DDC and NTIS do not tell authors how they should choose their "key words." Mr. Alex Hashovsky in his "Guide for Report Writers," AD 605443, has given the following concepts to be indexed:

Specific Materials, Data, Theories, Theses Used
Specific Properties Determined Experimentally or Theoretically
Specific Methods or Processes Investigated
Equipment Used
Specific Applications for Materials, Methods, Processes, or Equipment Where They Show Promise Beyond the Particular Experiment

These are stated to be applicable to any field. In many cases they should be restated in terms of the particular area where the file is of limited scope. In any case indexing which covers all the above concepts on which a document gives information and in the most specific form given in the text should make that item available for any search for which it is of value. Adjustments are desirable for special files. For instance the terms "alloy" and "plastic" apply to so many documents in a materials collection that they are of little if any value. Their use in a file on health would be little enough to make them of probable value. It should also be noted that every specific is part of a more general category and can therefore be indexed legitimately to the more general term. That a general term can be indexed to its subdivisions is NOT true. Information may be given only on class characteristics and none on its specific subelements.

How or by whom should the indexing be done? Although several publishers and organizations such as the Defense Documentation Center are requesting their authors to provide key words, the results are rarely satisfactory and further indexing is done by indexers employed by the filing organization. Such indexers are essential unless there has been at least some author indexing. Many feel that a computer can be programmed to index. It can be programmed to eliminate a large list of "common" words (conjunctions, prepositions etc., as well as those words which have no meaningful concept of their own, such as "part"). It can count the number of times a word or stem is used and even lump synonyms which have been indicated in its memory. However, determining whether information is given on the terms is another problem. Various groups are trying to develop instructions to meet this need but with very limited success so far. The author certainly knows what he has supplied information on and most authors want others to know and use their work. It would seem that authors who understand what good indexing is and know that only through good indexing can their work have maximum effective exposure should be both the best and the cheapest indexers.

There is a type of computer indexing that may be considered "quick and dirty" but that has considerable value. It is known as KWIC, Key Word In-Context, indexing (Fig 1). Titles are input to the computer with their access numbers, "common" or "stop" words eliminated, and the remaining terms arranged alphabetically accompanied with the rest of

REPORTS RECEIVED BUT NOT INDEXED

NOT READILY AVAILABLE DURING PROCESSING.
PLEASE REQUEST DOCUMENTS BY ACCESSION NUMBERS.

UIL	COOLER MADE OF TUBES WITH LOW SPIRAL FINS .*	87344
INTERNATIONAL	COOPERATIVE RESEARCH PROGRAM ON TOOL WEAR .*	27961
NICATION AND TEST OF HIGH-TEMPERATURE HONEYCOMB	CORE SANDWICH CONSTRUCTION .* FAB	41769
LA THE POLARIZATION EFFECT ASSOCIATED WITH THE	CORROSION FATIGUE OF V-95 ALLOY IN A SOLUTION OF A.NS	87288
ALE FOR THE ESTIMATION OF ORGANIC SUBSTANCES AS	CORROSION INHIBITORS .* MEANING OF THE ZETA-POTENTIAL	87348
MIXTURE OF QUINOLINE AND ISOQUINOLINE BASES AS	CORROSION INHIBITOR .*	87349
	CORROSION OF STAINLESS STEELS IN NAPHTHENIC ACIDS .*	87327
CYCLIC HOT SALT STRESS	CORROSION OF TITANIUM ALLOYS .*	40790
EN NIOBIUM AND TITANIUM .* INVESTIGATION OF THE	CORROSION RESISTANCE AND ELECTROCHEMICAL AND MECHANICA	87270
CRYSTALLIZATION OF	CORUNDUM FROM CRYOLITE MELTS .*	87301
L ELEMENTS PART E/ THE VARIATION IN FABRICATION	COST WITH SOME FUEL DESIGN PARAMETERS .* FABRICATION C	28212
WITH SOME FUEL DESIGN PARAMETERS .* FABRICATION	COSTS FOR PLUTONIUM FUEL ELEMENTS PART E/ THE VARIATIO	28213
NEUTRON IRRADIATION EFFECTS IN FCC, BCC, AND	CPH METALS .*	42801
POLYURETHANE ELASTOMERIC COMPOUND	CPS 746-BC, TYPE III. QUALIFICATION TEST REPORT .*	43092
NFLUENCE OF GEOMETRY ON THE STRENGTH OF FATIGUE	CRACKED PANELS .* A STUDY OF THE I	42793
HOT	CRACKING OF ALLOYS OF THE AL-CU-LI-MN SYSTEM .*	87280
FINE SLIPPING DURING	CREEP .*	87302
DEVELOPMENT OF NEUTRON SPECTROMETRY FOR	CRITICAL ASSEMBLIES .*	42896
PREPARATION OF	CROSS-LINKED POLYAMIDE FIBERS .*	87273
RS AND FLOW METERS .* CLOSED-SYSTEM TRANSFER OF	CRUDE OIL AND GAS PRODUCTION AND SOME NEW TYPES OF SEP	87304
CRYSTALLIZATION OF CORUNDUM FROM	CRYOLITE MELTS .*	87301
MAGNETOSTRICTION IN SINGLE	CRYSTAL TERBIUM .*	43080
	CRYSTALLIZATION OF CORUNDUM FROM CRYOLITE MELTS .*	87301
MANUFACTURING SINGLE	CRYSTALS .*	87343
PRODUCTION OF YTTRIUM IRON GARNET SINGLE	CRYSTALS IN THE DYNAMIC MODE .*	87279
PTICAL INVESTIGATIONS OF BARIUM TITANATE SINGLE	CRYSTALS IN THE INFRARED FREQUENCY RANGE .* DIELECTRIC	87370
THE ETCHED SURFACE IN SINGLE	CRYSTALS OF ANTIMONY .*	87256
ON THE MECHANISM OF GROWTH OF SINGLE	CRYSTALS OF CUPROUS OXIDE .*	87289
COLLOIDAL COAGULATION OF F-CENTERS IN KBR	CRYSTALS WITH IMPURITIES .*	87373
NT UPON THE STRUCTURAL PROPERTIES OF CAST STEEL	CR18NI9TI .* INFLUENCE OF HEAT TR	
N MELTING .* CAPACITY OF THE	CUPOLA AND THE PROCESSES OF COKE COMBUSTION AND C	
N THE MECHANISM OF GROWTH OF SINGLE CRYSTALS OF	CUPROUS OXIDE .*	
EMPIRATION OF THE SHRINKAGE OF ADHESIVES DURING	CURE .*	
ON PLASTIC INTERACTION	CURVES .*	
DEVELOPMENT OF IMPROVED	CUTTING FLUIDS AND LUBRICANTS .*	
INTERFACIAL POLYMERIZATION OF N-ALKYL ALPHA	CYANOACRYLATE MONOLOGS .*	
	CYCLIC HOT SALT STRESS CORROSION OF TITANIUM ALL	
TORSIONAL OSCILLATIONS OF AN ENCASED HOLLOW	CYLINDER OF FINITE LENGTH .*	42812
PARTICULATE	DAMPING IN SOLID PROPELLANT COMBUSTION INSTABILITY .*	43088
ASMA USING INDUCTION HEATING OF COLD PLASMA AND	DC PLASMA ENHANCEMENT .* PRODUCTION OF SUPERHEATED HYD	43101
ETIC MATERIALS OF GAMMA-FE2O3 TYPE, OBTAINED BY	DECOMPOSITION OF FERROUS OXALATE WITH COBALT IMPURITIE	87274
CATALYST OXIDES .*	DECOMPOSITION OF METHANOL ON MIXED ZNO-CRSUB 2 O SUB 3	87310
THE KINETICS OF THERMAL	DECOMPOSITION OF SILVER OXALATE .*	43110
ENCE OF TEMPERATURE ON THE MECHANISM OF PLASTIC	DEFORMATION OF MAGNESIUM AND MAGNESIUM ALLOY CONTAININ	87352
* MEASUREMENT OF ION AND ELECTRON	DENSITIES OF ELECTRON- BOMBARDMENT ION-THRUSTOR BEAM .	43097
LOW-ENERGY ACCELERATED ION	DEPOSITION .*	42807
OF PRECIOUS METALS .* EFFECT OF	DEPOSITION RATE AND ANNEALING ON THE OPTICAL CONTENTS	87390
INING THE DISTRIBUTION OF THE ACTIVITY FROM THE	DEPTH OF ALPHA EMITTERS .* DEVELOPMENT OF SPECTROMETRI	87294
LEX COMPOUNDS III. SILVER COMPLEXES OF PYRIDINE	DERIVATIVES IN ETHANOL-WATER SOLUTIONS .* THE EFFECT O	87257
* SYNTHESIS OF ESTERS AND OTHER CARBOXYLIC ACID	DERIVATIVES UNDER CONDITIONS OF ACID CATALYSIS FROM CA	87269

FOR OFFICIAL USE ONLY

9

Fig 1 Key-Word-In-Context

the title and the access number. Probably the most common and convenient form places the indexed terms at the left margin followed by the title in its normal sequence with the access number at the right. This is specifically called KWOC, Key-Word-Out-of-Context. This is excellent for rapid dissemination of new material but is not adequate for an information bank since it is an incomplete index, has virtually no control over the vocabulary, and includes many nonsignificant terms.

Index Storage

Indexing is the first step, is essential, and is expensive. Good indexing can be stored in any of the current files [illegible] each means of storage has both good and bad aspects. The type chosen should take into account the total environment for the system and optimize cost effectiveness.

[illegible]ost simple equipment for storage consists of blank cards, proba[illegible] 5" x 8", and a file drawer. The concept is written or typed at the top, usually left corner of each card and the concepts filed alphabetically. Access numbers of the documents so indexed are organized on the card below the heading. Frequently "terminal digit posting" is used. There are ten columns on the card, each containing only access numbers ending in one of our ten digits. The numbers in each column are in ascending order (Fig 2). This simplifies the manual comparison of access number lists by a factor of approximately 10. However, as the file grows manual searching becomes more difficult. Also manual handling of complicated questions especially involving the "or" relation are easy to err on even in a moderate sized file. This is still probably the best approach for many small private files.

The Peek-a-Boo, Light Coincidence, or Batten (from its initiator) system uses cards with the term or concept printed on a tab at the top (Fig 3). The card has an array of positions, usually 100 x 100, each of which identifies a particular document. There is no limit on the size of the vocabulary. "And" logic is easily handled, "or" is difficult or tedious except where hierarchical posting precoordinates the "or" group, and a set of negative cards is required for "not" logic. Equipment having a rather wide range of prices is required to drill an accurately positioned hole in the term cards to which a document is indexed. The cards are filed alphabetically and those for the terms of a search are pulled, superimposed, and viewed toward a light. The documents whose positions transmit light have been indexed by all desired terms. For moderate size collections unless the problems with "or" and "not" are serious this can be very effective.

Edge-notched cards (Fig 4) are used in many places. They have a distinct advantage in that they can have the abstract and/or vital information from the document on the central portion of the card so

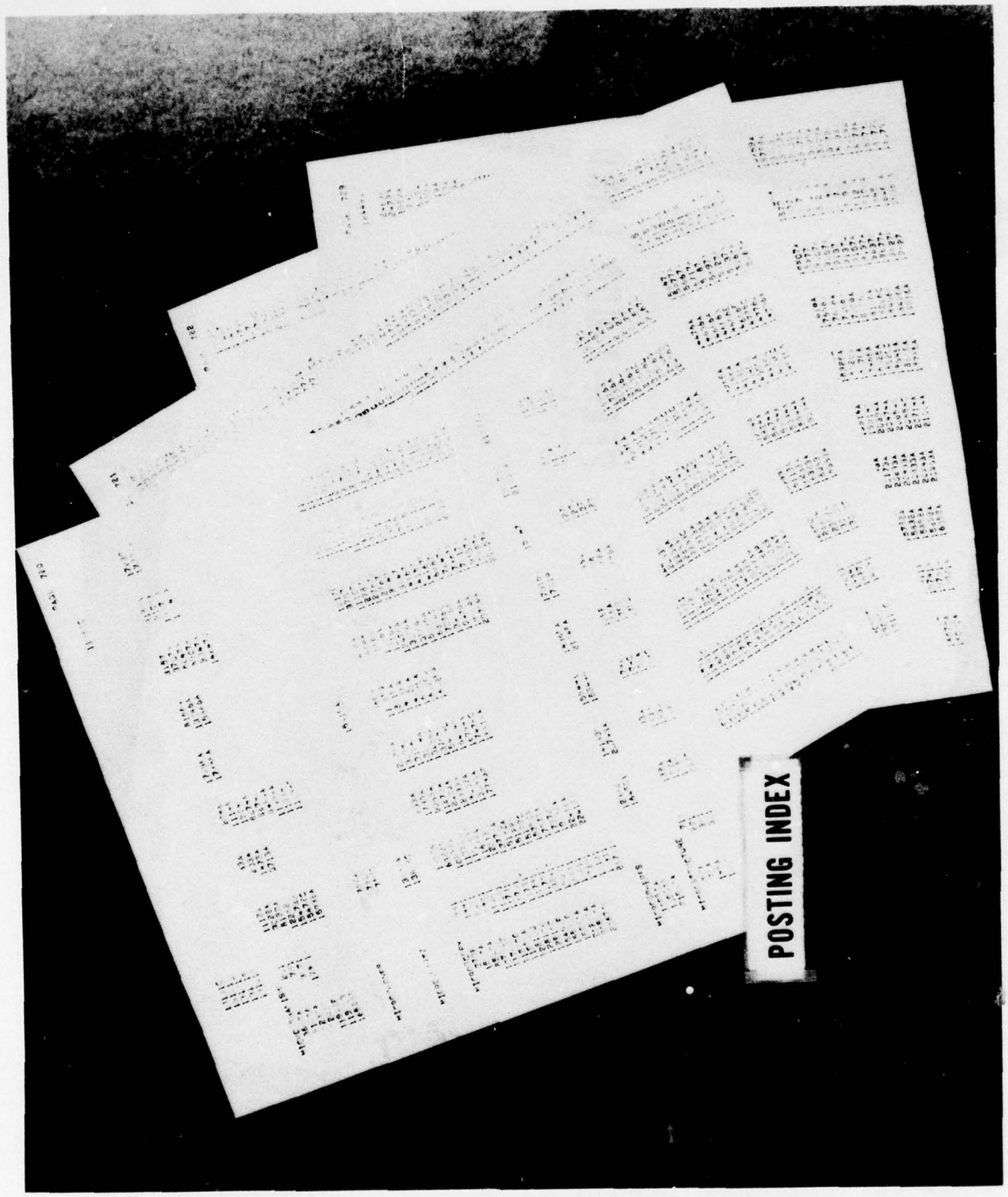


Fig 2 Index With Terminal Digit Posting

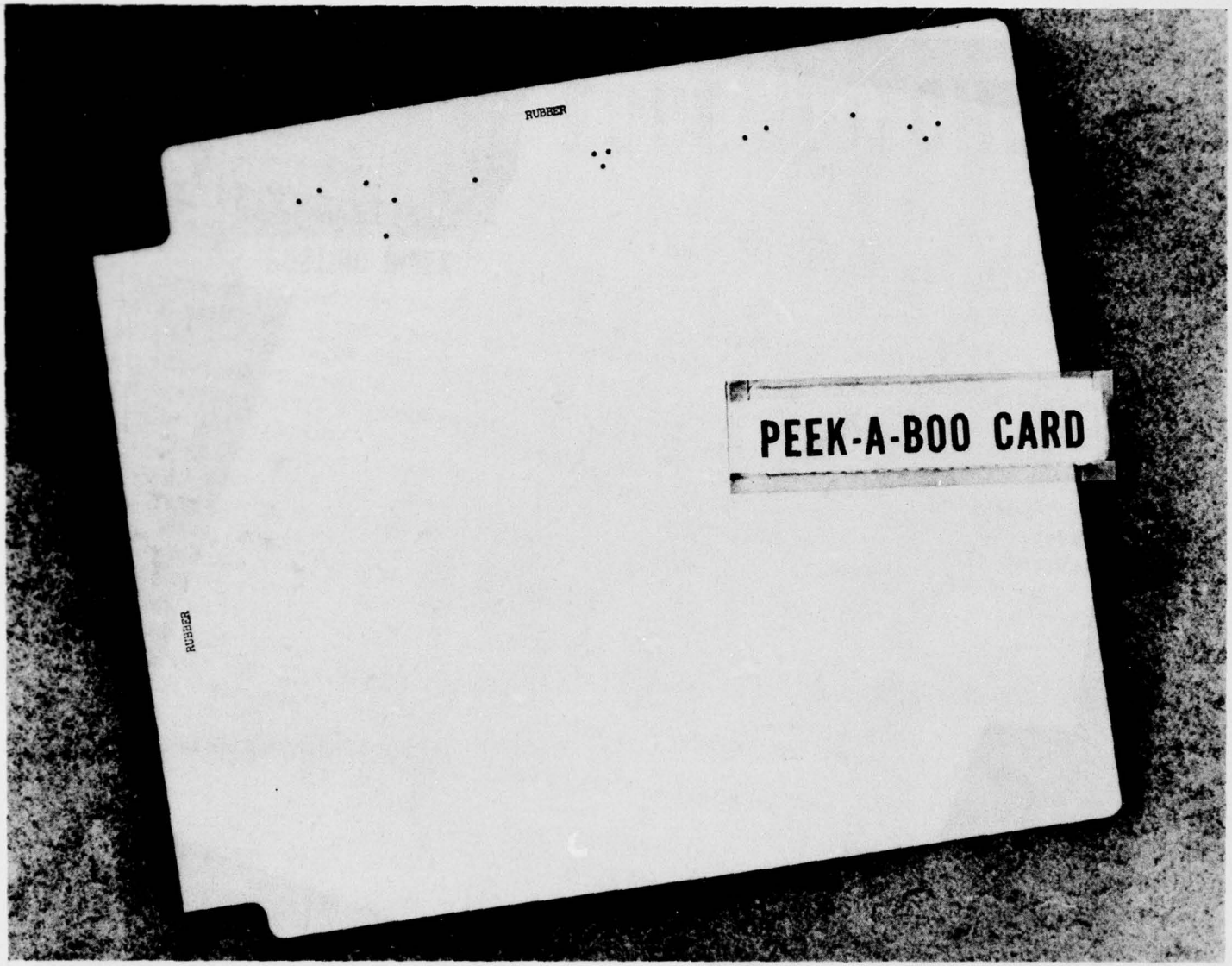


Fig 3 Light Coincidence Index

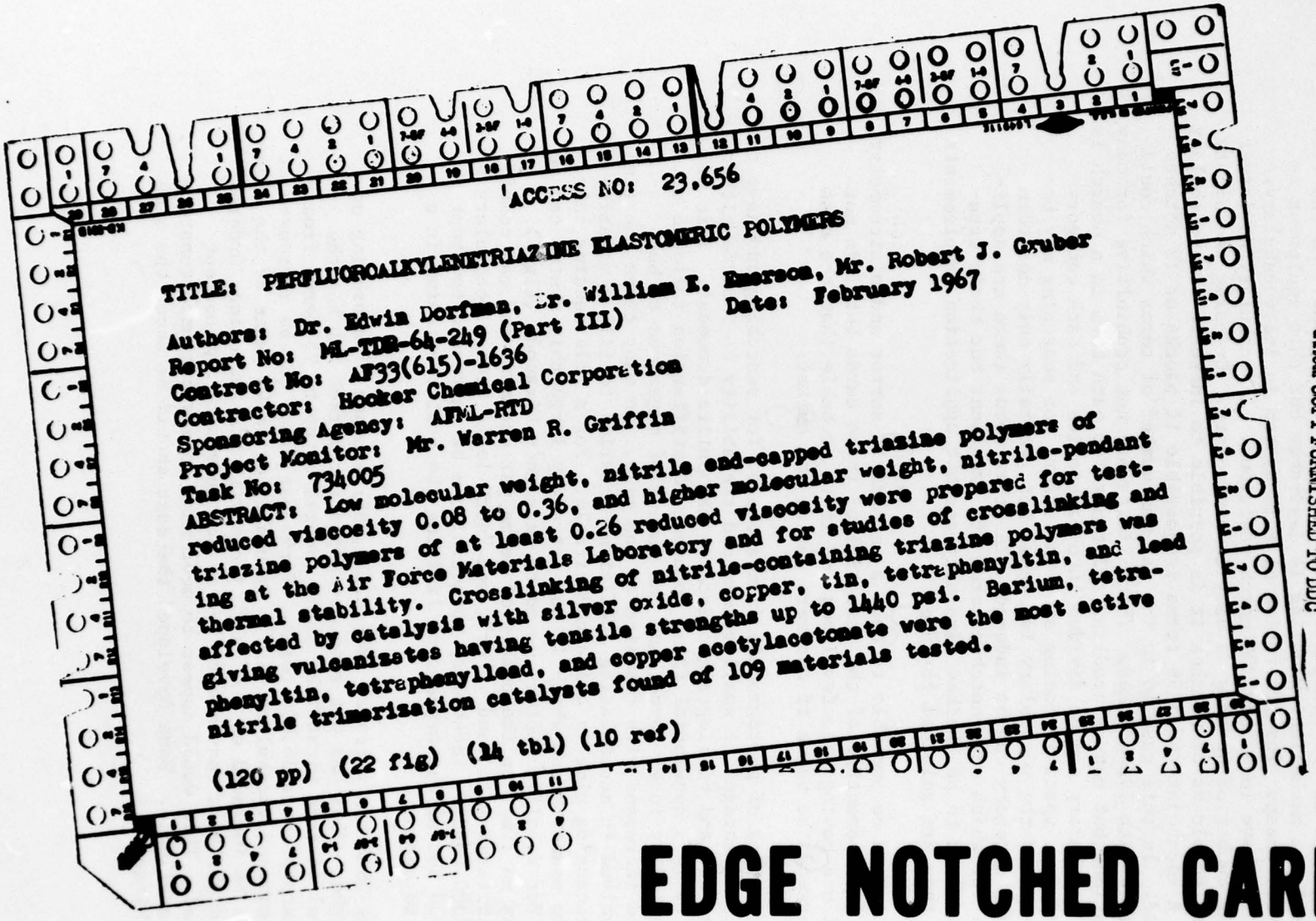


Fig 4 Edge Notched Card

that often there is no need to go to any other source. The cards are inexpensive as is the equipment to notch and sort by hand. As the file becomes large mechanical handling is desirable but this equipment is relatively cheap. The limitation of this system is its vocabulary. Most cards have less than 200 notch positions. A vocabulary no larger than the number of notches is easy to manipulate. By using the binary approach to blocks of notches it is possible to increase the vocabulary size. A vocabulary of 4096 terms is possible if blocks of 12 notches are used. In this case 16 is the maximum number of terms which could be indexed with 192 notches. These limits are not prohibitive for many applications but manual needling 192 times for each term in a search is. If the vocabulary can be divided into categories and each category assigned to a specific coding area the effort for searching may be decreased and the vocabulary increased but generally only one term from each category can be indexed. In other words there are applications for which edge-notched cards are excellent but their capabilities need to be checked carefully against application requirements before they are selected for use.

It would be possible to have a mechanical sorter and minicomputer to take full advantage of the potential of these cards but I do not know of one having been developed and it is probable that the cost would exceed the value if one were put on the market.

Magazines of photographic film are used for searching and have the distinct advantages of small volume and the ability to include either the abstract and bibliographic data or the entire document. This of necessity is a sequential file, document identification followed or accompanied by index terms, and the number of magazines to be searched increases with the size of the file. In many cases the time lag involved in accumulating sufficient material to fill a magazine plus processing time is quite undesirable. For a file needing only periodic updating and which is not extremely large this system can be used. For specific factors we go to Eastman's Miracode (Fig 5). This has no limit on the number of terms per document and can combine up to 15 terms with "and", "or", and "not" logic from a vocabulary of up to 2000 words. Vocabulary potential can be increased to about 4,000,000 but at the expense of limiting the number of terms in a search to seven.

The USAF has patented (No. 3,515,886) a means of searching on microfiche. This has not been developed for market but has the potential of each microfiche being indexed by up to 25 terms from a vocabulary of over 16,000 terms. Terms per fiche can be increased at the expense of vocabulary size. Naturally up to 98 pages of the document can be included on a 4" x 6" fiche with x24 reduction. Ordered filing might be a convenience but is not necessary for concept retrieval. This would appear to meet many needs but unfortunately is not available. When developed the cost should be about the same as for Miracode.

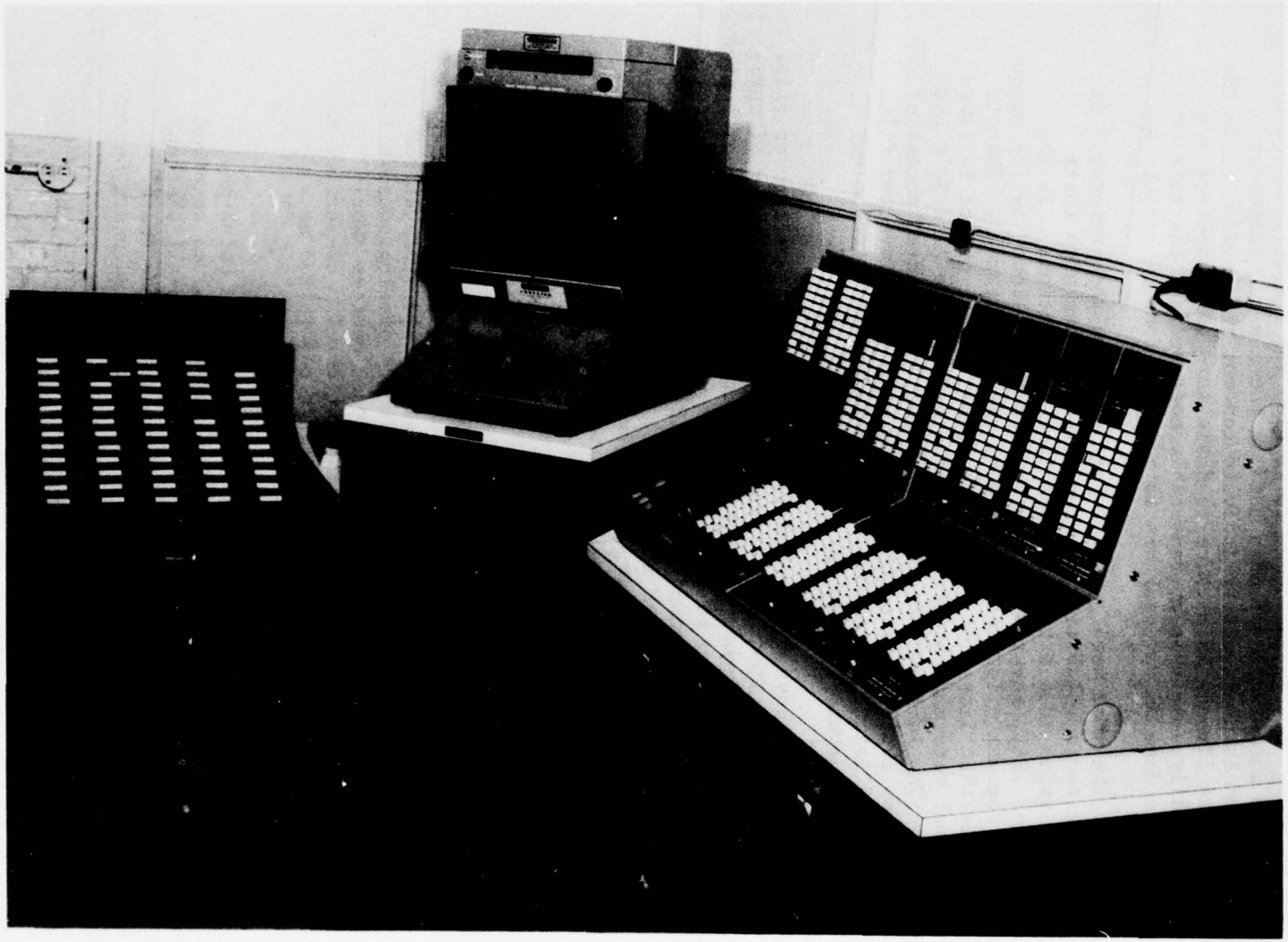


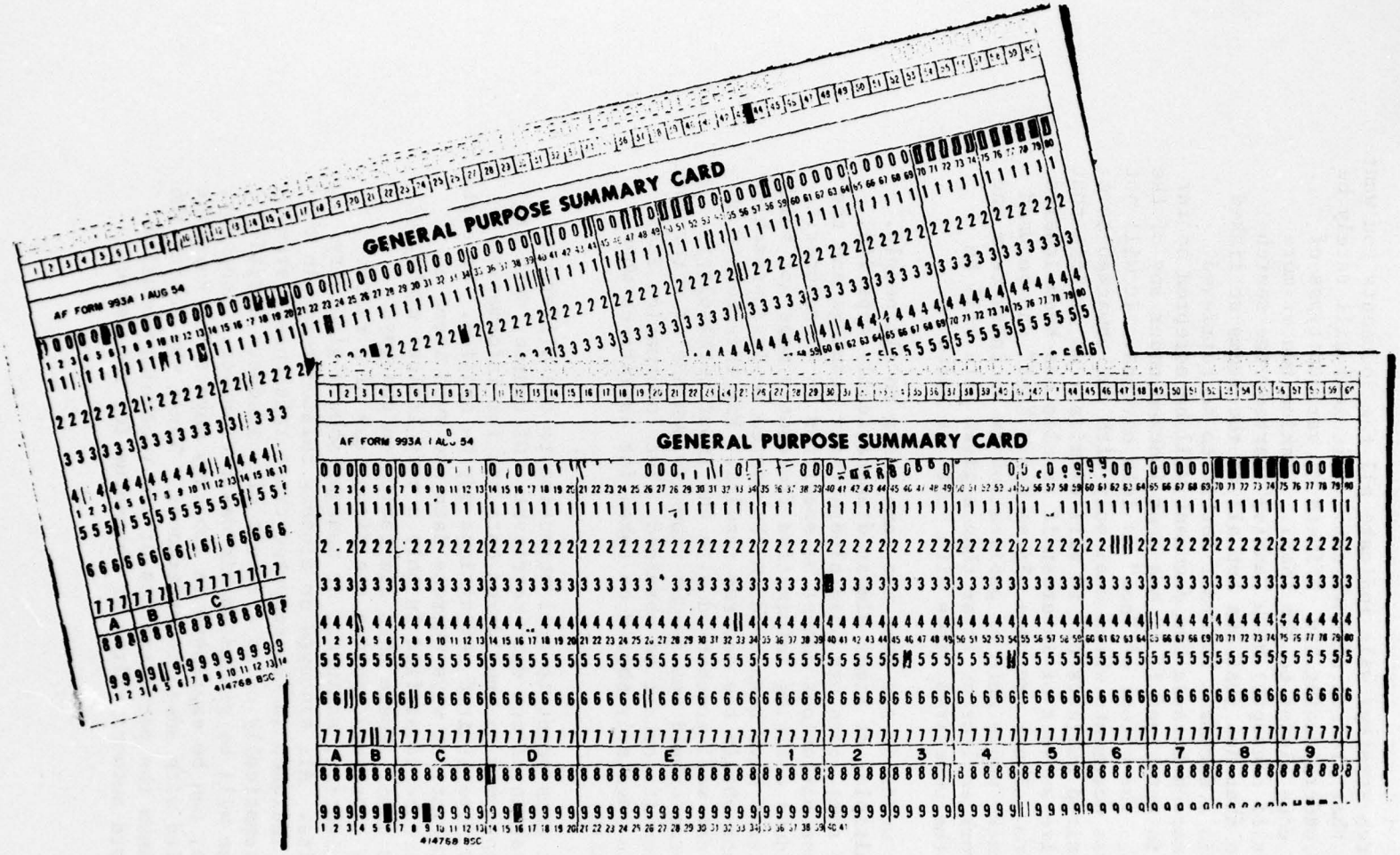
Fig 5 Miracode at Wright-Patterson Air Force Base

A system that has been and still is used some consists of punched EAM cards (Fig 6) and a card sorter. For sorting convenience each column or group of columns needs to be used for a specified set of values. This works well for data such as the numeric values of properties but is difficult to apply to concept retrieval. Since the information is already in machine-readable form and computers are becoming readily available at reasonable cost much of the retrieval that could be done on card sorters is actually done on computers.

Finally the most used index storage means is the computer. As stated above the availability of computers combined with their versatility makes them attractive for most applications. In the early days of IS&R, inverted files (The concept followed by the access numbers of documents indexed by it) were almost universal. Then it was realized that linear or sequential files (Access number followed by its terms) were more practical on computers of the late 50's and early 60's. The increased core capacity and the "random" access of disk files now available seem to indicate that again inverted files may be more efficient. Promised advances in computer technology indicate that further changes may be expected.

As indicated above a computer can manipulate either type of file. There is no absolute limit on the number of documents, number of terms per document, or the number of terms per question. All types of logic are available, "ands" can be included even within "or" groups etc. Any desired data can be included with the document identification, data may be manipulated after it is found, and searching can be done either batch-mode (off-line) or on-line from remote terminals. No wonder the computer is popular. However, to use these capabilities requires computer time and computer time is expensive. Searching and manipulating are usually well worth the cost. Using the computer as a printer for text is usually unduly expensive. At present most systems find it better to store text as paper or microfiche copy. There are microfiche retrieval units which can be linked to a computer and present the user with the microfiche documents selected by the computer. Programming can vary the efficiency of various operations over a considerable range and future developments will change the cost of some functions. Therefore, choices should optimize the operation for current capabilities with conversion to take advantage of future improvements always recognized as a possibility.

Probably all systems can operate off-line or in batch-mode. Speed of response and convenience are the primary values received from on-line operation. Many present terminals have all the capability of the computer so the only disadvantage is the cost. However, the use of remote terminals will be discussed under Search Strategy below.



MACHINE CARD

Fig 6 EAM Cards

Search Strategy

The ideal search strategy will indicate all the documents you want to review and none that are not of interest. This goal will rarely be reached. However, understanding the effects of various types of logic can help one come closer to the ideal. Linking two or more concepts (or groups of concepts) with "and(s)" narrows the search because it requires that not just one but all of the items so linked have been indexed in a document in order for it to be retrieved. "Or" broadens the search because the document will be accepted so far as the "or" group is concerned if it has been indexed under any of the concepts so linked. The effect of "not" is less obvious. It will not only eliminate those documents which deal only with the unwanted term and some of the desired subjects but it will eliminate an article that deals with the entire subject of interest if it also has been indexed under the "not" term. There are some places where "not" can be used safely. For instance, where date of publication is an index term and one is sure that work earlier or later than a given date will be valueless linking the undesired dates with "not" is safe.

Unfortunately words do not mean the same thing to all people. Some words have multiple, at times almost diametrically opposite, meanings. The way some words are used alters their meanings and the details of all these variations cannot be specified in a computer. Because the individual desiring information frequently does not fully know how certain terms have been used in the system he is approaching his question may not obtain the desired results. For this reason an individual who is not well acquainted with the input terminology practices of a system should not assume from his own use of a terminal that the desired results cannot be obtained. The Information Staff which uses the file regularly should be asked for assistance when there is any question.

It is well at this point to call attention to the difference between concept retrieval and data retrieval. Of course, data are concepts and concepts may be considered data but for information retrieval purposes more limited definitions may be applied. Concepts may be expressed in various ways whereas data are such items as properties which can be identified in only a very limited number of ways or a person's name. Where this limitation on the ways information can be expressed applies the direct use of a remote terminal by the scientist, engineer, or manager should give very satisfactory results. All synonyms or closely related terms in the vocabulary can be included in the computer cross reference systems so that either automatically or as a result of a feedback notification the searcher will be guided to his desired answer. Concepts, as specified above, can be expressed in too many ways to be sure the computer is supplied with adequate references. Those who search the system regularly know the peculiarities of its vocabulary-usually have the system word authority list or thesaurus for reference.

A good thesaurus can be very helpful even if it was not designed for or used by that particular system. Thesaurii (Fig 7) as used in IS&R have concepts listed alphabetically and give broader, narrower, and related terms. BT, NT, and RT are not the only symbols used to indicate these concepts. Some Thesaurii have "scope notes" which clarify the term's used in that system. Regardless of the expressions used the relations are the same and can guide one to terms which may have been used in another system. The array of concepts related in various ways to the initial term can be used to help one decide precisely what he needs when his initial concepts are vague.

Document Storage and Retrieval

We have indicated that some retrieval mechanisms provide only access numbers and some can provide the abstract or even the whole document. Rarely is directly retrieving the entire text practical if it exceeds a page or two; often not even then. So the documents must be stored for convenient access. Each one must have a distinct number. Therefore, it is customary for each item simply to be assigned the next consecutive number as it is accessioned into the system. The document, its abstract, and its indexing all carry this same access number. Where reports come from various sources the source may be indicated by letters preceeding the number. The National Technical Information Service keeps DDC's access number with its initial letters, AD, for DDC reports. Many other accessions carry the previously assigned number to avoid confusion. Obviously each organization has to choose its own method of assigning access numbers. A simple system is desirable and having more than one access number so that cross-referencing is required is generally to be avoided.

Actual storage of paper copy may be in anything from simple shelving to elaborate, power driven files. Microfiche and other document forms may be kept on any convenient storage medium. The abstracts also may be so stored if they are not part of the index file. In any case they should be ordered logically for convenience in locating desired documents. As mentioned earlier future developments in magnetic and/or optical storage are likely to make some changes in the most practical storage means. Larger systems may be expected to keep abreast of these developments and managers of smaller files should try to keep informed either directly or through the major centers.

Output Requirements

Regardless of what the initial product of the search may be the requester wants information. It may be data, abstracts or documents, but a high portion of the material he gets should be closely related to his request.

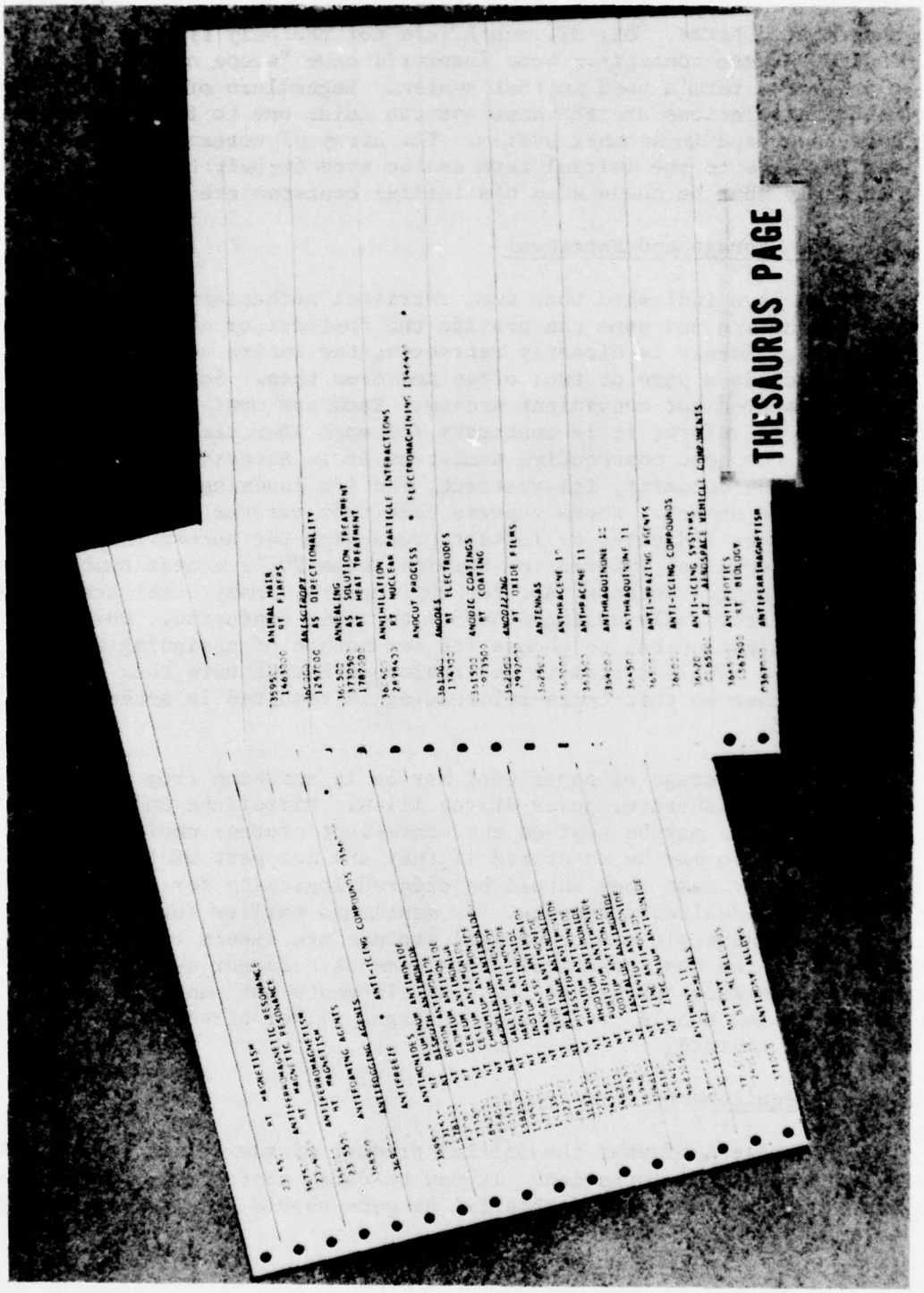


Fig 7 Sample Thesaurus Pages

If data are to be the end product there are several factors to be considered. Is the need for a single point, a table, or plotted data? What accuracy is required and does the source indicate the probable quality of the data either directly or indirectly? Should the data be for certain circumstances such as temperature, humidity, etc.? Is the data with its probable accuracy all that is desired or should the source be indicated so that further detail can be checked as desired? Exactly what is desired may not be available, but a clear understanding of both the desire and the capability of the system will improve customer satisfaction. Some systems can manipulate or compute data. This makes interpolation and extrapolation possible if needed. Also if this capability is available the user should know the possibilities in case they are of value to him.

If his interest is in what we have labeled conceptual information a clear understanding of the subject of interest is the first requirement. Sometimes browsing is necessary for the individual to clarify his own thinking. A review of the terms related to his interest in a thesaurus is an excellent first step and may well crystalize the requirement. If this is inadequate a few somewhat related documents or their abstracts may be quite helpful. What is the breadth or specifity of the need and is a moderate amount of information on the subject all that is desired or is all the available information wanted? As with data systems, the exact need may be impossible to satisfy in the file or files within the system. Recognition of this fact before the search can modify the users anticipation and create a much better reaction.

As indicated earlier the proper use of "and", "or", and "not" can tailor the question and therefore the response to the desire. Some systems have additional means for improving selectivity. Weighting is used by several systems. Here each term is assigned a numerical value and the computer does not give a response unless the sum of the values of terms used for its indexing equals or exceeds the value assigned for that search. This prevents certain irrelevant documents from being retrieved but at times a number of low value terms add up to the threshold without any high value term having been used. Some have used weighting on the input but the interests of the searcher may differ greatly from those of the author and cause input weighting to be misleading.

Another approach which has had very little publicity but deserves definite consideration is used in a search system at Wright-Patterson Air Force Base. Programming for this file permits the question to be framed with the term or terms that are essential placed first. Others follow in order of decreasing importance. The computer prints first those access numbers which meet the maximum number of requirements and then goes back step at a time to the "cut off" point which includes

only the essential items (Fig 8). With the print-out of this type of search one first reviews those references printed first and proceeds only till his desires are satisfied - or he concludes this approach is fruitless.

The above techniques are used largely in "batch-mode" however, there is no reason they couldn't be used "on-line." The favored method when using remote terminals seems to be for the computer to indicate the number of references responsive to each question and when this number satisfies the searcher he reviews the list of documents. To a great extent the terminal user is naturally using the principles applied to "batch-mode" in that he broadens or narrows his search till the retrieval is about the size he wants. As he examines these references he can revise or completely reframe his search if need be to be sure he gets from the file any of its contents which will help him.

Feedback

Feedback or customer response is extremely important for several reasons. It may frequently indicate inadequacies in the holdings, and in cases where all users agree, unnecessary acquisitions. If adequately detailed it can give clues as to how to improve indexing and searching. The use being made of the output may well assist in determining the most cost effective approach under various circumstances. Most significant from many points of view is information on value received. Management wants to know the effectiveness of its expenditures. If the return on investment is really good, financing will probably continue and possibly increase. Users will have the service and the information staff will have jobs. No matter how much value is actually realized, if management is not made aware of this value both user and staff are likely to suffer.

Means of getting and recording feedback will vary. For technical detail face to face communication is usually best but in large organizations not totally practical. Questionaires are frequently used and provide written evidence of the worth. Most of the time the dollar value of information is difficult to establish. Since dollars are management's measure it is of utmost importance to give dollar values and their basis wherever possible. Unmeasureable values should also be given with reasonable detail. They are not likely to receive the same consideration as dollars but they are recognized.

INFORMATION RETRIEVAL -- DOCUMENT SEARCH NO 12677 DATE 7 APR 72

SEARCH TITLE - WELDBONDING AEROSPACE STRUCTUR REQUESTED BY - T GUZIK NASA/LEWIS CUTOFF 9 MAX NO. AL.
LAB. NO. IS R-3204

ORDER	CONN	WORD NO.	
1	AND	58000	ADHESIVES
2	OR	57300	ADHESIVE BONDING
3	AND	4573000	WELDING
4	OR	008500	BRAZING
5	OR	4563000	WELDS
6	AND	71500	AIRCRAFT
7	OR	65500	AEROSPACE VEHICLE COMPONENTS
8	OR	3894700	STRUCTURAL COMPONENTS
9	OR	3736500	SPACE VEHICLES
10	AND	57300	ADHESIVE BONDING
11	AND	4273000	WELDING
12	AND	3894700	STRUCTURAL COMPONENTS
13	AND	71500	AIRCRAFT
14	OR	65500	AEROSPACE VEHICLE COMPONENTS

PROBABLY PERTINENT
X REQUIRES EVALUATION
PROBABLY NOT PERTINENT
SEND TO: T. GUZIK

17

038322A 065789A

THE 2 DOCUMENTS LISTED ABOVE CONTAIN THE FIRST 14 WORDS.

054377A 041463A C07207A

THE 3 DOCUMENTS LISTED ABOVE CONTAIN THE FIRST 12 WORDS.

027380B 029504A 040923A C30659A 052457B

THE 5 DOCUMENTS LISTED ABOVE CONTAIN THE FIRST 11 WORDS.

026413A 043047A 058774A C59572A 007306A

THE 5 DOCUMENTS LISTED ABOVE CONTAIN THE FIRST 10 WORDS.

032405A 065637A C08938A

THE 3 DOCUMENTS LISTED ABOVE CONTAIN THE FIRST 9 WORDS.

Fig 8 Wright-Patterson Air Force Base Print-Out

A factor for consideration primarily by the information staff is "insurance." Duplicates of documents received from large repositories are replaceable again from the same source and that source usually retains some type of back-up. The magnetic records as received from commercial sources are similarly protected. However, any records locally produced or altered should be considered from the standpoint of what their loss would mean. Fire and "acts of God" can destroy any record. A back-up duplicate of printed matter, possibly microform, kept in a different location may be well worth its cost. Magnetic records face an additional hazard, demagnetization either by accident or gradual deterioration. Punched update cards and a tape or disk that has been updated can constitute a good back-up and the storage space they require is small. Naturally they should not be kept too close to the record being used since a single catastrophe might destroy both.

Information Sources for Unusual or Special Needs

It would be nice if the local system could meet all local needs promptly. Even to strive for this ideal would be impractical since it is impossible to predict the borderline and seemingly unrelated records that will actually be wanted. The efficient approach is to know where virtually any item may be obtained if needed in a reasonable time and at reasonable cost.

Probably the most complete sources for commercially published material of all types are the Library of Congress, New York Public Library, and the John Crerar Library in Chicago. However, many books, periodicals, etc. are available from near-by public and college libraries. These are usually readily available on inter-library loan. Also a "Union List of Periodicals" and other cross-reference tools are in daily use at these facilities. Which of these tools a particular library should keep itself depends on usage. However, any library should have some means of obtaining their use.

There are also many highly specialized information and/or data services. Most of these were initially financed by the Government but have recently been requested or required to provide at least 1/2 of their funding by selling their products. On the surface this arrangement seems quite reasonable but many individuals and companies who will bet on a horse race are unwilling to pay for an information file search which may simply show that no work has been done in the area of interest. This circumstance causes changes in these centers from time to time. The "Directory of Federally Supported Information Analysis Centers" is updated periodically by the Committee on Scientific and Technical Information and is available through the National Technical Information Service. All editions to date have included more than 100 such sources of information. The reliability

of out put varies from a simple collection of values that have been published in a given discipline to the highly evaluated products of the Thermophysical Properties Research Center which have received the approval of the National Bureau of Standards. Many fields of endeavor are so related that a center in the general field of interest can direct one to the best source available for a specific question.

Available output from these services seems to include the complete gamut of the possible. Defense Documentation Center, National Technical Information Service, and others supply actual documents in hard copy or microfiche, abstracts with bibliographic data, or magnetic records that can be searched by the user. However, referenced material that has been copyrighted must frequently be obtained from a commercial vendor. Pandex, Engineering Index, Chemical Abstracts and others provide search media, (magnetic tape, microfiche, abstract journals, etc.) giving bibliographic information but no documents. Data centers usually provide just that -- data. But the sources of the data are also generally included should more detail be needed.

The means of obtaining these services are as varied as the output. Visits, letters, wires, and phone calls are all used. However, in a world of rising costs some costs are decreasing and having a profound influence on the handling of information. The cost of virtually all computer functions has reduced drastically and that of cross-country telecommunications has dropped distinctly. Further downward changes, at least for service rendered, are anticipated. These facts make the use of remote terminals highly attractive and the trend is toward more and more on-line service. The values and limitations of terminal use have already been discussed. As remote linkage to computer data bases becomes the cheapest approach to information systems some of the problems will be reduced if not eliminated. However, one should always use the various helps available before assuming that a system will not provide satisfactory results.

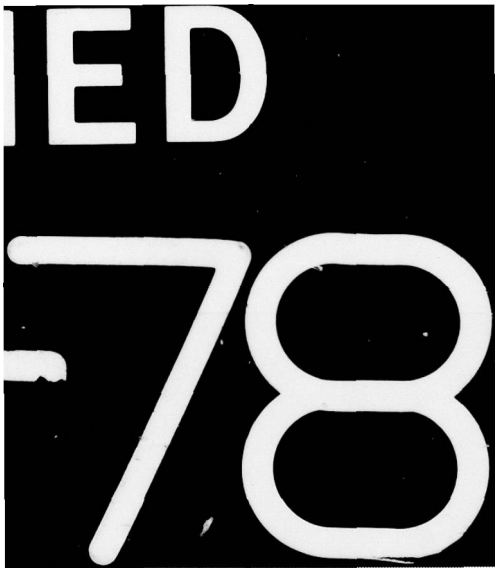
There are two basic types of terminals in use. With both the input is by means of a typewriter type keyboard. One type of unit types both request and answer on paper and, thus, provides a permanent record of the intercommunication. The other has a cathode ray tube on which input and output are both displayed. Previous communication must be removed to make room for current recording. Many of these units include the capability of printing what is on the screen if desired and/or having the desired portions printed at the computer and mailed. The variety of available terminals is great as a review of the ads in any issue of Datamation will make obvious. If one is procurring equipment he should check the attributes of several units relative to the needs of his system and computer file requirements before choosing one.

The user of an operating system will do well to check its manuals and talk with other users or the designer of the system to make sure he knows how to use it to the best advantage.

With most present systems the data bank cannot be altered from the terminal. The equipment in many cases could be given this ability and providing the opportunity to add material is being seriously considered if not actually practiced in several cases. Probably the worst effect adding can have is to clutter the file and require purging at a later date. Altering the stored record is quite a different matter. Either with the best of intentions or maliciously a user could make a record completely violate the intent of the author. Cases of computer fraud have had wide publicity where remote access was not involved. Even with all the partial safeguards that have been developed the potential for either accidental or intentional damaging of a valuable information store is too great generally to permit the users to make changes from a terminal.

A further convenience in input and output is being studied - voice communication. Some success has been achieved but the problems of different voices, different pronunciation, and homonyms will certainly delay the general application of vocal communication with the computer.

In conclusion, whether one is setting up a storage and retrieval system for a group of users, applying a present system for others, or seeking help on a individual question the first requirement is to establish what is wanted. It is well also to determine what can reasonably be expected so that the results won't be disappointing.


ED
78