A Selective Current-Awareness System Using Engineering Index's Plastics Data Base. I. System Design*

R. H. WAGNER
Research Laboratories, Eastman Kodak Co., Rochester, N. Y. 14650

T. H. HIGGINS
Management Systems Development Department, Kodak Park
Division, Eastman Kodak Co., Rochester, N. Y. 14650

Received October 1, 1968

Details are presented on the design and operation of a selective dissemination system predicated on the use of a descriptorized data base. The elements found to be of most value in the operation of the system are described in detail. Interest profiles are constructed on the basis of grouping logic, which enables the user and the system manager to prepare and to maintain profiles more easily; various supplemental reports are supplied to the user, some regularly and some occasionally, to aid in the improvement of profile effectiveness. The format of the notification card enables the user to deduce much about an article's content without reference to it or to the abstract, and it provides a useful means to instruct the system manager in the event that profile adjustment is required.

As a part of a general program being conducted by the Department of Information Services of the Kodak Research Laboratories, an arrangement was made in April 1966 with Engineering Index, Inc., to use their plastics literature file as the basis for the development of a selective current-awareness system that could be administered easily and maintained with sufficient effectiveness to justify expanding it. Engineering Index, Inc., 345 East 47th St., New York, N. Y. 10017, through their User's Participation Program, will supply all interested people with bibliographic material and other services related to its use.

The system became operative in August 1966 with eight individuals as clients, or users. The clientele was expanded to 20 by adding two new clients each month during the following six-month period.

The performance of the system is covered in the following paper.

THE DATA BASE

At the present time, Engineering Index has two machinable bases: a Plastics Section and an Electrical/Electronics Section. Although we are working with both, the data in this and the following paper are limited to the Plastics base. The operating system described here is compatible with both bases.

The format of Engineering Index's machinable data bases comprises five segments: title (or notation of content); author(s); source reference; list of descriptors, with associated subdescriptors; and subject headings. Each segment of each article references an accession number, which indicates the location of an indicative abstract available in nonmachinable form.

Material for this data base is drawn by Engineering Index from a group of 375 journals and miscellaneous publications, including the U. S. Patent Gazette. Abstracts are published by them in a monthly bulletin, i.e., the *Plastics Section*, which at present comprises about 500 items each month.

Descriptors and subdescriptors are drawn from an authority list (thesaurus) comprising approximately 16,000 descriptors and 40 subdescriptors, and are assigned by the professional staff at Engineering Index. The number of descriptors assigned varies from about 6 to 30 per article; in the majority of cases the number will fall between 10 and 16.

THE OPERATING SYSTEM

The system operates on a minimum configuration of an IBM 1460, 16K Computer with four tape drives, an IBM 1403 Printer, and a card reader-punch. The operating system and its related output will be described with reference to Figures 1 through 8.

Figure 1 shows a simplified flow diagram of the first part of the system. The bibliographic, or master file, material received from Engineering Index is extracted by a program (SD-1000) which produces a document segment tape and a document segment report. The document descriptors are extracted from the segment tape, sorted into an inverted listing arrangement, and input into the matching program (SD-2007). This and other "SD-Programs" were written expressly for this project. These

* Presented before the Division of Chemical Literature, 155th Meeting, ACS, San Francisco, Calif., April 1968.

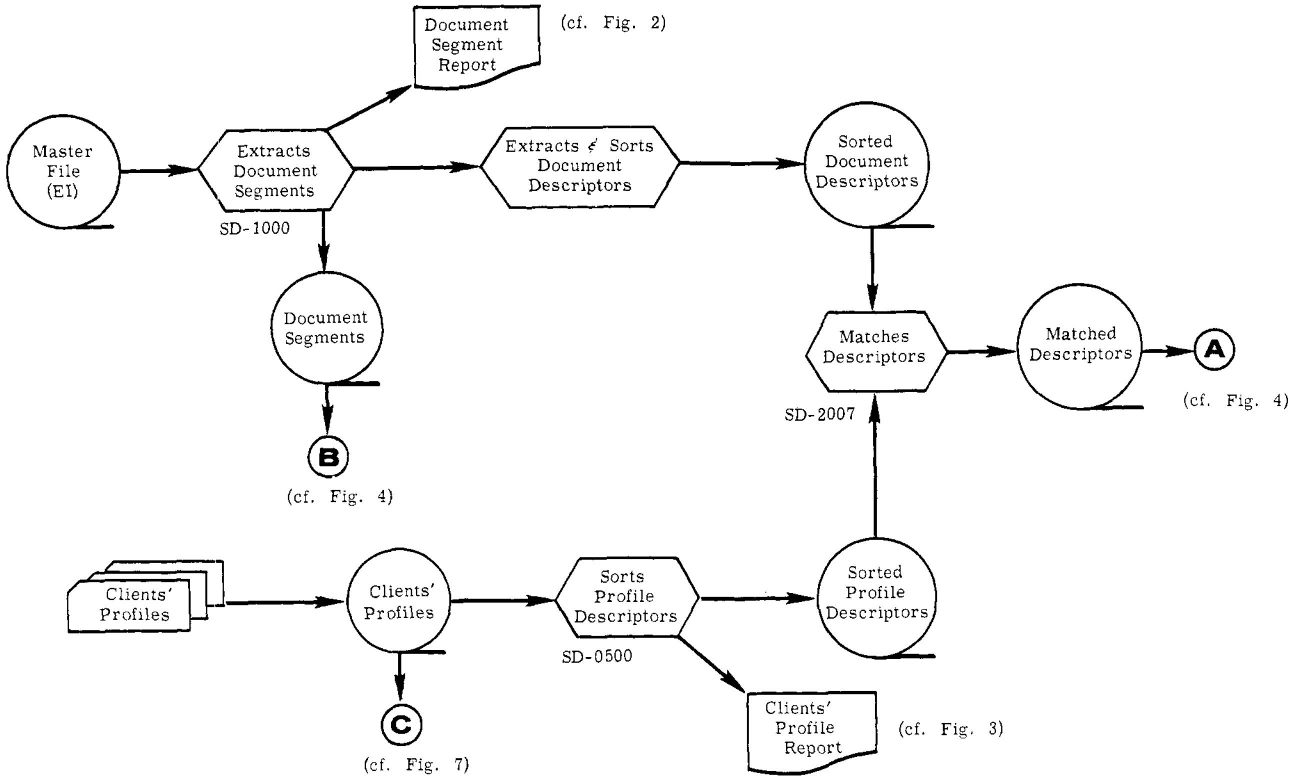


Figure 1. Operating system - part I

programs are supplemented by other "utility programs" which were immediately available, e.g., various sort and summary programs. All programs were written in Multicoder—a programming language used internally at Eastman Kodak Co.

The clients' interest profile cards are converted to tape (each run) and input into a special sort and print program (SD-0500) which arranges the profile descriptors into an inverted listing order for input to SD-2007 and also prints the clients' profile report.

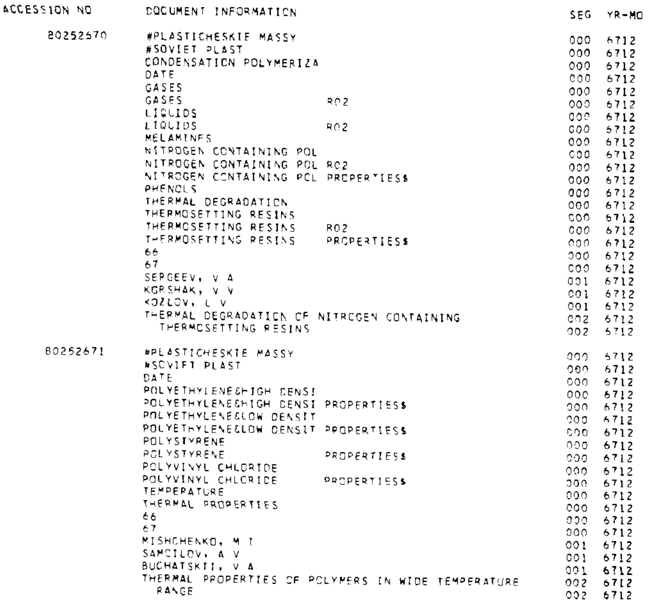
The document segment report, shown in Figure 2, displays the contents in each of the segments of each article in the master file. This report is used for internal control.

The clients' profile report, shown in Figure 3, is sent to each client each month. It lists, for each subprofile, the descriptors and subdescriptors comprising each parameter (indicated by the control value) and suggests, by its title, the interest area the subprofile is intended to cover.

The matching program (SD-2007) compares the document descriptors against the interest-profile descriptors, and whenever a match occurs the descriptor (or the descriptor-subdescriptor) term is output onto a matched-descriptor tape; each output term is associated with one or more document pointers and with one or more subprofile pointers, together with their associated control values.

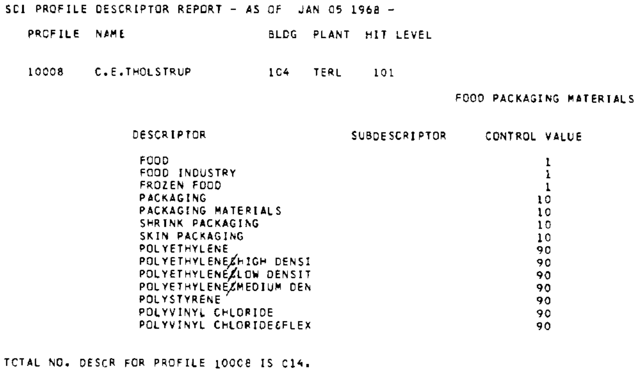
This material is then sorted to bring all matches for a given document and given subprofile together and in this form is input into a validation program (SD-3002) (Figure 4).

The validation program determines if the required logic, dictated by the control values, is met; this will be described in somewhat greater detail later. This program also suppresses any duplication of references that might occur by the selection operation of two or more subprofiles. When the same document is selected by several subprofiles through different combinations of matching descriptors, the notice that ultimately emerges will appear to have been selected by the subprofile having the lowest number.



```
ACCESSION NO    DOCUMENT INFORMATION                                          SEG  YR-MO

  80252670      #PLASTICHESKIE MASSY                                          000  6712
                #SOVIET PLAST                                                 000  6712
                CONDENSATION POLYMERIZA                                       000  6712
                DATE                                                          000  6712
                GASES                                                         000  6712
                GASES                      R02                                000  6712
                LIQLIDS                                                       000  6712
                LIQUIDS                    R02                                000  6712
                MELAMINES                                                     000  6712
                NITROGEN CONTAINING POL                                       000  6712
                NITROGEN CONTAINING POL R02                                   000  6712
                NITROGEN CONTAINING POL PROPERTIES$                           000  6712
                PHENOLS                                                       000  6712
                THERMAL DEGRADATION                                           000  6712
                THERMOSETTING RESINS                                          000  6712
                THERMOSETTING RESINS       R02                                000  6712
                THERMOSETTING RESINS       PROPERTIES$                        000  6712
                66                                                            000  6712
                67                                                            000  6712
                SERGEEV, V A                                                  001  6712
                KORSHAK, V V                                                  001  6712
                KOZLOV, L V                                                   001  6712
                THERMAL DEGRADATION OF NITROGEN CONTAINING                    002  6712
                  THERMOSETTING RESINS                                        002  6712

  80252671      #PLASTICHESKIE MASSY                                          000  6712
                #SOVIET PLAST                                                 000  6712
                DATE                                                          000  6712
                POLYETHYLENE&HIGH DENSI                                       000  6712
                POLYETHYLENE&HIGH DENSI PROPERTIES$                           000  6712
                POLYETHYLENE&LOW DENSIT                                       000  6712
                POLYETHYLENE&LOW DENSIT PROPERTIES$                           000  6712
                POLYSTYRENE                                                   000  6712
                POLYSTYRENE                PROPERTIES$                        000  6712
                POLYVINYL CHLORIDE                                            000  6712
                POLYVINYL CHLORIDE         PROPERTIES$                        000  6712
                TEMPERATURE                                                   000  6712
                THERMAL PROPERTIES                                            000  6712
                66                                                            000  6712
                67                                                            000  6712
                MISHCHENKO, M I                                               001  6712
                SAMOILOV, A V                                                 001  6712
                BUCHATSKII, V A                                               001  6712
                THERMAL PROPERTIES OF POLYMERS IN WIDE TEMPERATURE            002  6712
                  RANGE                                                       002  6712
```

Figure 2. Document segment report



```
SCI PROFILE DESCRIPTOR REPORT - AS OF  JAN 05 1968 -
   PROFILE  NAME                     BLDG  PLANT  HIT LEVEL

   10008    C.E.THOLSTRUP            104   TERL    101
                                                           FOOD PACKAGING MATERIALS

                DESCRIPTOR                    SUBDESCRIPTOR    CONTROL VALUE

                 FOOD                                                1
                 FOOD INDUSTRY                                       1
                 FROZEN FOOD                                         1
                 PACKAGING                                          10
                 PACKAGING MATERIALS                                10
                 SHRINK PACKAGING                                   10
                 SKIN PACKAGING                                     10
                 POLYETHYLENE                                       90
                 POLYETHYLENE&HIGH DENSI                            90
                 POLYETHYLENE&LOW DENSIT                            90
                 POLYETHYLENE&MEDIUM DEN                            90
                 POLYSTYRENE                                        90
                 POLYVINYL CHLORIDE                                 90
                 POLYVINYL CHLORIDE&FLEX                            90

TCTAL NO. DESCR FOR PROFILE 10008 IS C14.
```

Figure 3. Client's profile report

Program SD-3002 produces two tapes. One, labeled "Hit-Descriptors," is a record of all of the matched descriptors that have passed the logic requirements, together with their document and interest-profile pointers. The other, labeled "Hit-Documents," is a record of all of the accession numbers of the documents selected by the system. The latter tape is input to another extraction program (SD-1005), where it acts as a signal to cause this program to extract the segments of all of the required documents from the document segments tape (B of Figures 1 and 4). The output from SD-1005 interfaces with the integrating and print program (SD-3500), which combines the information on the hit-descriptor tape and prints the notification card shown in Figure 5. The details of this card will be described later.

Another element of the operating system is shown diagrammatically in Figure 4. This has to do with the production of a "New Descriptor" list and the updating of dictionary tapes on which records are maintained, giving the use frequencies of each item, as well as the information relating to the first and the last time that each term was used to index a document.

The new descriptor list, shown in Figure 6, is produced each month; a copy of it is sent to each client. It serves to alert the clients to useful items for updating profiles. Periodically, reports are printed from the dictionary tapes; they are used as a guide in the construction of new profiles and the amending of existing profiles. The format of this report is the same as that shown in Figure 6.

The remaining element of the operating system is shown in Figure 7, which illustrates the uses made of Engineering Index's thesaurus tapes. The master thesaurus tape is updated each month by the use of Engineering Index's thesaurus updating program, TH3, using as input the "Additions & Deletions" tape supplied by them. This program also prints a report of all recent changes in the authority list (thesaurus), which is used as the basis for making any necessary revisions in interest profiles.

When required, the updated master thesaurus tape can be extracted by Program SD-0099, producing a tape which contains only the main-entry ("Main") terms and another tape containing all entry terms plus all associated cross-reference terms. The first of these output tapes is used to validate all interest profiles occasionally, to be certain that no errors exist in spelling, spacing, or use of invalid terms. This check is made whenever a new profile is introduced into the system or when major revisions are made to existing profiles.

The other tape output from SD-0099 is used as the basis for producing subsets of the main thesaurus, comprising only that material which is relevant to existing profile descriptors; we call it the "Personalized Thesaurus-Extract." Programs SD-8000 and SD-8500 were written to extract only those thesaurus elements dictated by the

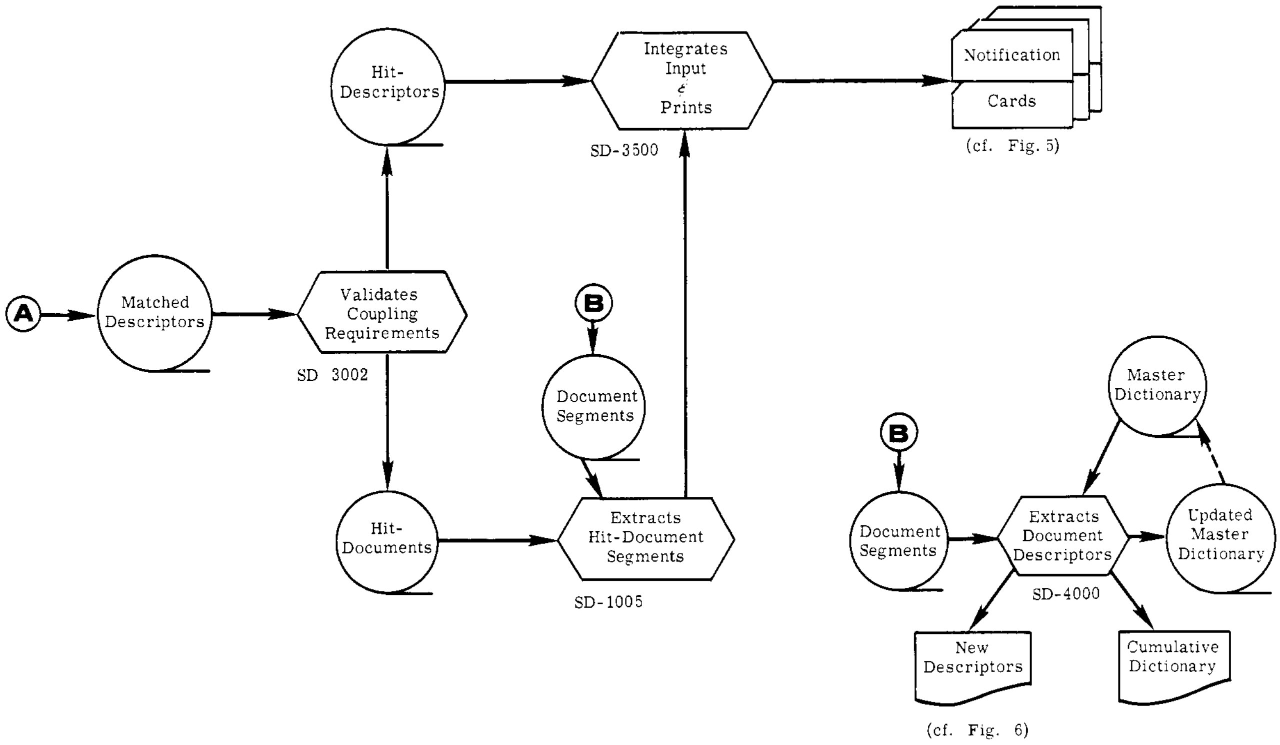
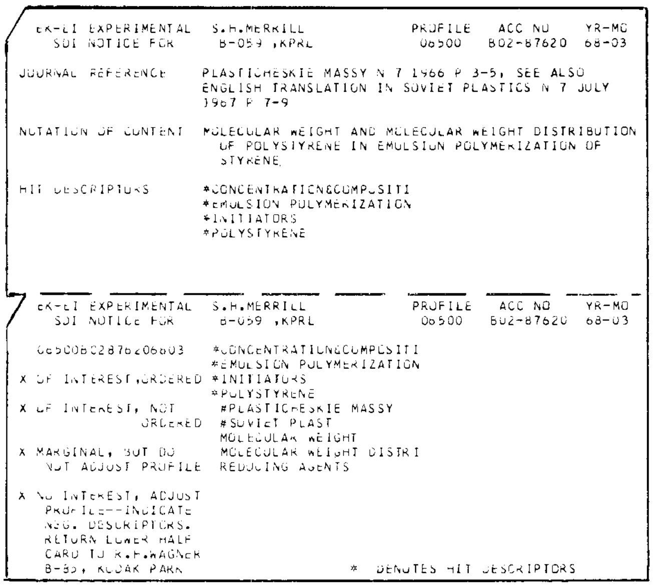
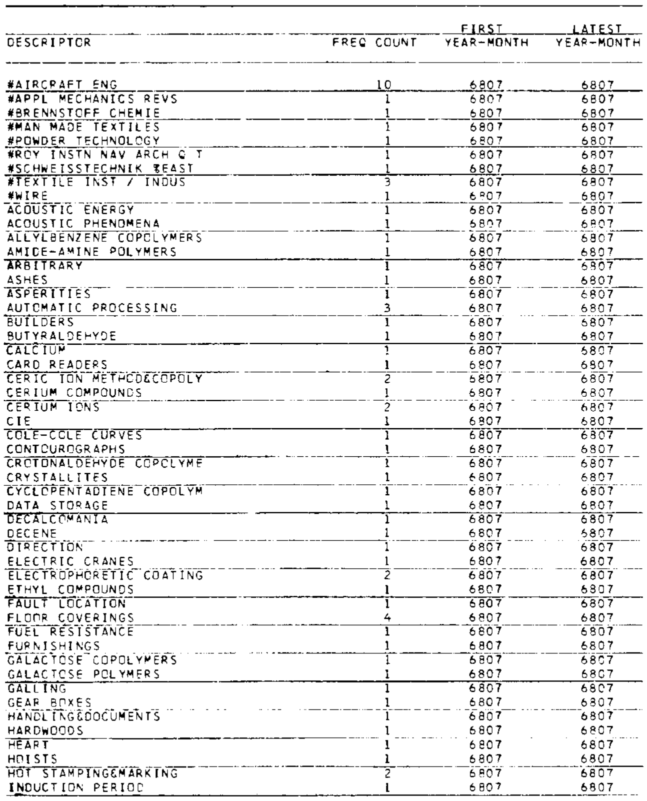


Figure 4. Operating system - part II



```
EK-EI EXPERIMENTAL  S.H.MERRILL              PROFILE   ACC NO     YR-MO
  SDI NOTICE FOR     B-059 ,KPRL              06500   B02-87620  68-03

JOURNAL REFERENCE   PLASTICHESKIE MASSY N 7 1966 P 3-5, SEE ALSO
                    ENGLISH TRANSLATION IN SOVIET PLASTICS N 7 JULY
                    1967 P 7-9

NOTATION OF CONTENT MOLECULAR WEIGHT AND MOLECULAR WEIGHT DISTRIBUTION
                      OF POLYSTYRENE IN EMULSION POLYMERIZATION OF
                      STYRENE

HIT DESCRIPTORS     *CONCENTRATICN&COMPOSITI
                    *EMULSION POLYMERIZATION
                    *INITIATORS
                    *POLYSTYRENE

EK-EI EXPERIMENTAL  S.H.MERRILL              PROFILE   ACC NO     YR-MO
  SDI NOTICE FOR     B-059 ,KPRL              06500   B02-87620  68-03

  06500B0287620 6803  *CONCENTRATIONGCOMPOSITI
                      *EMULSION POLYMERIZATION
X OF INTEREST,ORDERED *INITIATORS
                      *POLYSTYRENE
X OF INTEREST, NOT     #PLASTICHESKIE MASSY
              ORDERED  #SOVIET PLAST
                       MOLECULAR WEIGHT
X MARGINAL, BUT DO     MOLECULAR WEIGHT DISTRI
   NOT ADJUST PROFILE  REDUCING AGENTS

X NO INTEREST, ADJUST
   PROFILE---INDICATE
   NEG. DESCRIPTORS.
   RETURN LOWER HALF
   CARD TO R.F.WAGNER
   B-65, KODAK PARK                       *  DENOTES HIT DESCRIPTORS
```

Figure 5. Notification card



DESCRIPTOR	FREQ COUNT	FIRST YEAR-MONTH	LATEST YEAR-MONTH
#AIRCRAFT ENG	10	6807	6807
#APPL MECHANICS REVS	1	6807	6807
#BRENNSTOFF CHEMIE	1	6807	6807
#MAN MADE TEXTILES	1	6807	6807
#POWDER TECHNOLOGY	1	6807	6807
#ROY INSTN NAV ARCH Q T	1	6807	6807
#SCHWEISSTECHNIK %EAST	1	6807	6807
#TEXTILE INST / INDUS	3	6807	6807
#WIRE	1	6807	6807
ACOUSTIC ENERGY	1	6807	6807
ACOUSTIC PHENOMENA	1	6807	6807
ALLYLBENZENE COPOLYMERS	1	6807	6807
AMIDE-AMINE POLYMERS	1	6807	6807
ARBITRARY	1	6807	6807
ASHES	1	6807	6807
ASPERITIES	1	6807	6807
AUTOMATIC PROCESSING	3	6807	6807
BUILDERS	1	6807	6807
BUTYRALDEHYDE	1	6807	6807
CALCIUM	1	6807	6807
CARD READERS	1	6807	6807
CERIC ION METHOD&COPOLY	2	6807	6807
CERIUM COMPOUNDS	1	6807	6807
CERIUM IONS	2	6807	6807
CIE	1	6807	6807
COLE-COLE CURVES	1	6807	6807
CONTOUROGRAPHS	1	6807	6807
CROTONALDEHYDE COPOLYME	1	6807	6807
CRYSTALLITES	1	6807	6807
CYCLOPENTADIENE COPOLYM	1	6807	6807
DATA STORAGE	1	6807	6807
DECALCOMANIA	1	6807	6807
DECENE	1	6807	6807
DIRECTION	1	6807	6807
ELECTRIC CRANES	1	6807	6807
ELECTROPHORETIC COATING	2	6807	6807
ETHYL COMPOUNDS	1	6807	6807
FAULT LOCATION	1	6807	6807
FLOOR COVERINGS	4	6807	6807
FUEL RESISTANCE	1	6807	6807
FURNISHINGS	1	6807	6807
GALACTOSE COPOLYMERS	1	6807	6807
GALACTOSE POLYMERS	1	6807	6807
GALLING	1	6807	6807
GEAR BOXES	1	6807	6807
HANDLING&DOCUMENTS	1	6807	6807
HARDWOODS	1	6807	6807
HEART	1	6807	6807
HOISTS	1	6807	6807
HOT STAMPING&MARKING	2	6807	6807
INDUCTION PERIOD	1	6807	6807

Figure 6. New descriptor list

terms (descriptors) in those profiles specified by a set of control cards. A typical page of this report is shown in Figure 8; its use will be described below.

The entire system involves 27 programs, of which 13 were existing utility programs and the other 14 were written expressly for this system. Of these 14 programs, 12 are used in the main operating system; the other two are used as required for producing root-extract reports and term-term association reports. Their use will be described later.

INTEREST PROFILES. CARD FORMAT

Interest profile cards, representing either an individual or a project group, are characterized by a 3-digit numeric field in Cols. 1–3. Each individual (or group) usually has many subprofiles; these are designated by a 2-digit numeric field in Cols. 4, 5. The name of the individual (or group) occupies Cols. 6–26; the address, Cols. 28–35. The first four spaces of the address field are used to specify the building or department number, followed by a 4-letter Coden-like label to specify the plant, division, department, etc. Col. 36 is used for a code to specify the length of the match field; if blank, the match field will be limited to the DESCRIPTOR term only (Cols. 37–59); if an 11-punch (-), the match field will be extended to include the SUBDESCRIPTOR term (Cols. 60–71). The parameter code—sometimes referred to as the control value or control code—is a 2-digit signed numeric field in Cols. 72, 73. A separate card is required for each descriptor or descriptor-subdescriptor required, and each card must contain all of the above elements.

INTEREST PROFILES. LOGIC AND STRUCTURE

All profile elements, i.e., subprofiles, use unweighted Boolean logic. The descriptors characterizing the items of interest are arranged in OR-groups; the control value (CV) which is assigned to each descriptor determines its match criteria. If the match criterion is unity (absolute interest) for one or more descriptors, each is assigned a CV = 99+. If conjunction between two OR-groups of descriptors is desired, i.e., a match criterion of 2, each item of one group is assigned a CV = 11+ and each item of the other group is assigned a CV = 90+. If conjunction between three OR-groups is desired, i.e., a match criterion of 3, each item of one group is assigned a CV = 01+, the second group a CV = 10+, and the third group a CV = 90+; an example of this type of conjunction is shown in Figure 3. Negation control is provided by assigning CV = 99− to each negating descriptor; negation overrides all other logic in our system.

When conjunction logic is involved, the validation program (SD-3002) is designed to reject any match configuration that does not provide for at least one match in each of the conjoining parameters. For example, if matches are found between any number (up to 10) of (CV 11+) descriptors and no matches are found in any of the (CV 90+) descriptors, no notification card will issue; similarly if two, or more, matches are found in (CV 90+) descriptors but with none of the (CV 11+) descriptors, no notification will issue. The six control values are numeric codes that are used by the validation program (SD-3002) in determining the coupling requirements of the matched descriptors. In this program, all CV99+ and CV99− matches are converted to magnitudes of 101+ and 500−, respectively: all other CV-categories are retained at their assigned magnitudes. The program then sums the converted and the unconverted CV's for all matches found within *each* document profile and within *each* interest subprofile and compares this sum against the following three conditions which must be met if the

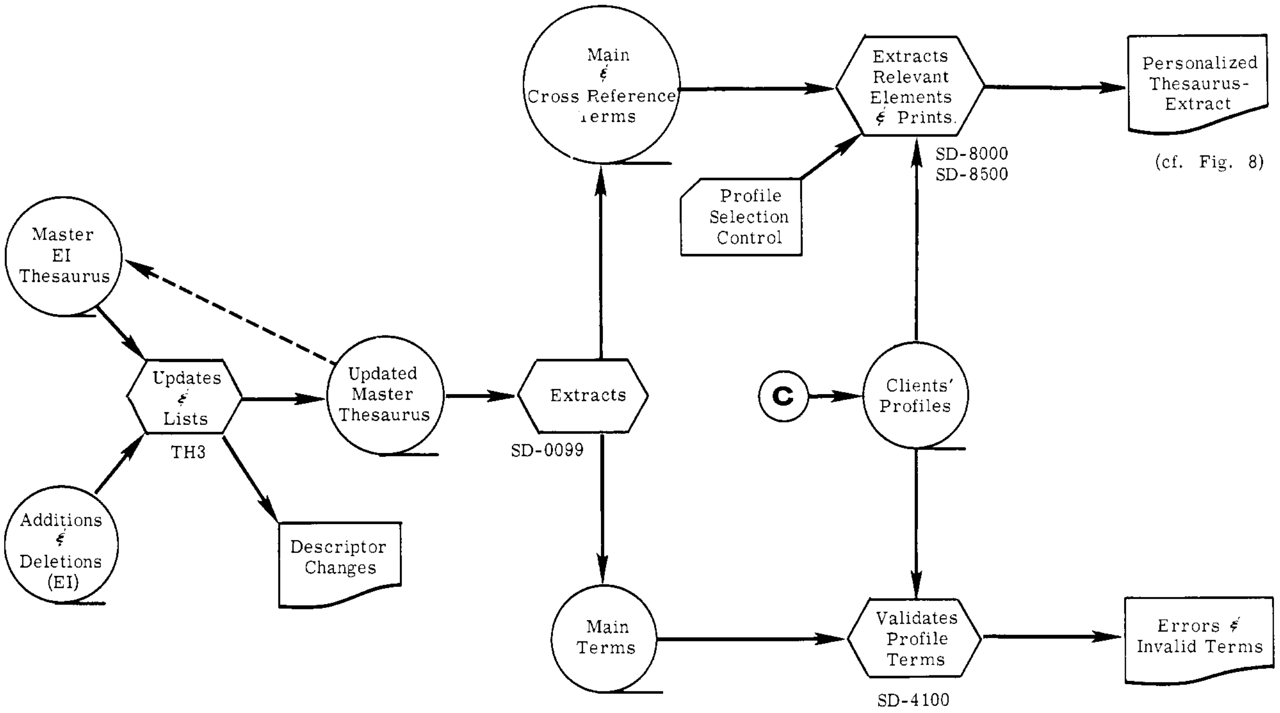
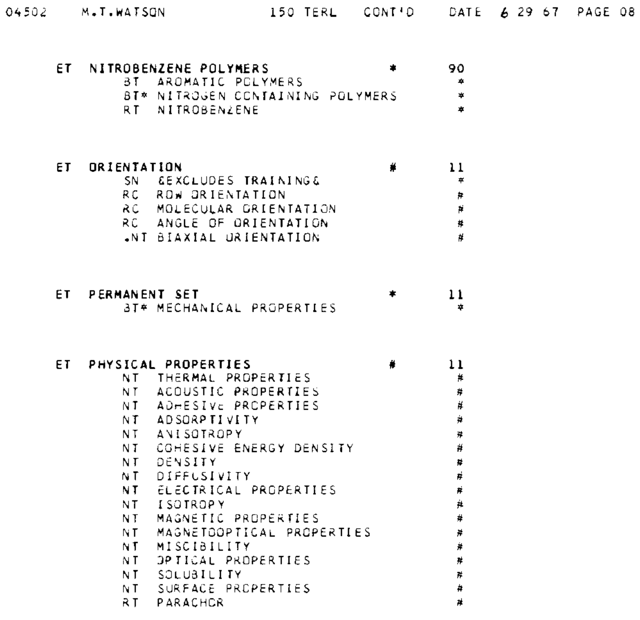


Figure 7. Operating system - part III



```
04502    M.T.WATSON          150 TERL   CONT'D    DATE  6 29 67  PAGE 08


     ET  NITROBENZENE POLYMERS              *       90
           BT  AROMATIC POLYMERS                     *
           BT* NITROGEN CONTAINING POLYMERS          *
           RT  NITROBENZENE                          *


     ET  ORIENTATION                        #       11
           SN  &EXCLUDES TRAINING&                   #
           RC  ROW ORIENTATION                       #
           RC  MOLECULAR ORIENTATION                 #
           RC  ANGLE OF ORIENTATION                  #
          .NT  BIAXIAL ORIENTATION                   #


     ET  PERMANENT SET                      *       11
           BT* MECHANICAL PROPERTIES                 *


     ET  PHYSICAL PROPERTIES                #       11
           NT  THERMAL PROPERTIES                    #
           NT  ACOUSTIC PROPERTIES                   #
           NT  ADHESIVE PROPERTIES                   #
           NT  ADSORPTIVITY                          #
           NT  ANISOTROPY                            #
           NT  COHESIVE ENERGY DENSITY               #
           NT  DENSITY                               #
           NT  DIFFUSIVITY                           #
           NT  ELECTRICAL PROPERTIES                 #
           NT  ISOTROPY                              #
           NT  MAGNETIC PROPERTIES                   #
           NT  MAGNETOOPTICAL PROPERTIES             #
           NT  MISCIBILITY                           #
           NT  OPTICAL PROPERTIES                    #
           NT  SOLUBILITY                            #
           NT  SURFACE PROPERTIES                    #
           RT  PARACHOR                              #
```

Figure 8. Personalized thesaurus extract report

coupling is to qualify as valid and a notification card is to issue:

1) The summed-CV must be equal to or greater than 101+ (the hit-level);

2) The summed-CV must be a number that does not end in zero;

3) The summed-CV, when right-truncated by one digit, must be a number that is not evenly divisible by 9.

If it is desirable to provide for a valid conjunction between items in any single parameter of a given subprofile, a separate subprofile can be set up to supplement the "parent subprofile."

THE NOTIFICATION CARD

The notification card, shown in Figure 5, is a vertically arranged double-card, both halves of which contain the client's name, address, profile, and subprofile number, document accession number, and the year-month of the abstract bulletin. The upper half contains, in addition, the full reference to the document, the notation of content (or title) of the document, and a listing of the "hit descriptors." The lower half of the card contains, in addition, a listing of all of the other descriptors assigned to the document, i.e., the complete document profile. Space is available for a listing of up to 30 descriptors by restricting the print field to 23 characters (which is identical to the match field used in the EK/EI system) and by omitting a listing of any associated subdescriptors. This latter is a definite disadvantage, for it not only deprives the client of valuable information about the document content, but also puts both the client and the system manager at a disadvantage when profile adjustments are required.

In the normal operation of the system, the client is asked to read the abstracts bearing the accession numbers on his notification cards in his advance copy of a complete collection of abstracts covering that month's output. He is also asked to rate each reference in one of four ways: (1) Of Interest, (copy of original document) Ordered, (2) Of Interest, (copy of original document) Not Ordered, (3) Marginal, Do Not Adjust (Sub) Profile, (4) Of No Interest.

In the event of a "4-rating," the client is asked to indicate any of the listed descriptors that best characterize the reason for this rating, together with any annotations that might be helpful in profile adjustments.

INTEREST PROFILE—CONSTRUCTION AND USE OF EXTRACT REPORTS

The construction of interest profiles was an experiment in itself. Various methods were tried during the period covered by this report, including selection by the client of descriptors pertinent to his interests drawn from a list of about 5,000 items, i.e., using a dictionary report. This method did not prove to be very popular, although it was definitely more acceptable than selecting terms from the bulky master thesaurus. The method recommended at present has been generally used since mid-1967 for profile construction; it involves the following steps: An initial collection of terms pertinent to written descriptions of the client's interests is assembled; this selection is made by the system manager using the master thesaurus, dictionary reports, etc. The match criteria (CV) of the various groups of terms are then determined by the manager and client. Where indicated, existing root extract reports are reviewed with the client to expand coverage. For example, in an interest involving "rubbers," the appropriate section in the root extract report where descriptors involving rubber materials are collected is shown to the client, who indicates additional terms to be included in his profile. After the prototype profile is keypunched for introduction into the operating system, a personalized theasurus extract is prepared for the client. This provides a display of cross-referenced terms for each of the terms in his prototype profile, from which he chooses additional descriptors to increase profile effectiveness.

The program developed for extracting all descriptors and all descriptor fragments containing any specified character-set, provides the option of fore-truncation, aft-truncation, fore- and aft-truncation, or a complete word. These programs also allow listing the extracted material in either a report display or in the form of punched cards in the required format for direct introduction into the operating system. Experience has shown that it is not economically practicable to use the punched-card option; it is preferable to produce reports from time to time relating to various root terms as needed, and to use these reports for general reference in the manner previously described.

Little use has been made of the term-term association report, which gives the frequency that any descriptor has been used in indexing with any other descriptor. Originally, it was believed that this kind of report would be very useful as a reference guide in profile construction; however, the cost of producing it, even on a limited number of documents and a limited number of specified entry-descriptors, was too high to enable us to justify a fuller exploration of its value.

A Selective Current-Awareness System Using Engineering Index's Plastics Data Base. II. Performance*

R. H. WAGNER
Research Laboratories, Eastman Kodak Co., Rochester, N.Y. 14650

Received October 1, 1968

The operational performance over a 17-month period of the previously described selective dissemination system is presented. Of the 21,000 notifications sent to about 20 users, 91% were evaluated; of these, 14% were of "Document-Ordered Interest," 48% were "Of Interest," 27% were "Marginal," and 11% were "Of No Interest." Recall data obtained from about half the users over a period of eight months show the precision-factor/recall-factor products are generally greater than 0.5. The effect of iterative profile adjustments on precision-recall performance is discussed. A comparison made with four other SDI systems shows a relatively high level of performance for this system.

During the 17 months of operation of the system (August, 1966, through December, 1967) covered in this paper, we have sent out 21,000 notifications to an average of 17.8 individual clients. We have received "relevance ratings" for about 19,100 of these (91%), which average out as follows:

Of Interest, Document Ordered -- 2750 references (14.4%)
Of Interest, Document Not Ordered -- 9070 references (47.6%)
Marginal -- 5100 references (26.8%)
Of No Interest -- 2140 references (11.2%)

Figure 1 shows the percentile ratings on a monthly basis.

In the first six months of operation [from August, 1966 (6608), through January, 1967 (6701)], clients were offered only three rating categories; beginning with the 6702 issue,

* Presented before the Division of Chemical Literature, 155th Meeting, ACS, San Francisco, Calif., April 1968.