Encoding Internal Reports at the Philip Morris Research Center*

By DONALD P. MURRILL
Philip Morris, Incorporated, Richmond, Virginia
Received April 30, 1962.

The system used for encoding internal reports at the Philip Morris Research Center of Philip Morris, Incorporated, virtually insures that only those reports which are pertinent to a question asked of the system are retrieved. The use of IBM cards punched with primary and secondary keywords and their modifiers and coded to show the relationship between the keywords permits rapid retrieval of the pertinent documents.

The Philip Morris system is a variation of the collating method of indexing, in which an IBM card is assigned to each keyword. It was developed after we had experimented with the scanning system. Initially, it was thought that the most desirable feature of a coding system would be that all the coding concerning a given report could be punched into no more than one or two IBM cards. It was for this reason that the scanning system was investigated first. The assigned report number and general information concerning the subject matter were punched in code into the first few columns of the card. Keywords were punched in code in the remaining columns.

Figure 1 shows the document card for Report Number 430 (columns 4–6), which was written in December, 1952 (columns 16–18). The information contained in Report Number 430 was described by the following keywords:

32–9	Tobacco
33–2	Latakia
33–7	Turkish
48–0	Moisture meter
58–2	Moisture content
68–6	Calibration

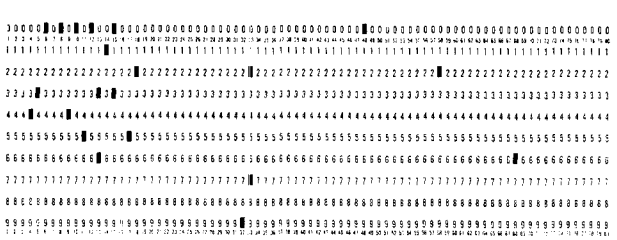


Fig. 1.

An intelligent guess about the subject matter of Report Number 430, on the basis of these words, would be that it concerned Turkish and Latakia tobaccos and the calibration of a moisture meter for measuring moisture content. Or, perhaps, it concerned the moisture content of Turkish and Latakia tobaccos and the calibration of a moisture meter. The true relationships cannot be determined from the keywords alone.

As the number of keywords used to describe the content of a report increased—and most reports required considerably more than six keywords—the relationships became correspondingly more obscure.

A hierarchical system was used for setting up the dictionary for the scanning system. The following illustrates the arrangement:

23–0	Agronomy
32–9	Agronomy-Tobacco
25–9	Agronomy-Tobacco-Production
34–X	Agronomy-Tobacco-Production-Topping and Suckering

The number in the left column pertained to the keyword furthest to the right on the horizontal line and showed the position of that word on the IBM card. When a keyword was used for describing the information in a document, all of the words to the left of that keyword were also used. Thus, the information was described on levels ranging from generic to specific. For example, a report which discussed the topping and suckering of tobacco had listed, besides topping and suckering, agronomy, tobacco, and production.

To make it possible to find any word in the dictionary and its position code, each word was listed in the first column of words, and this column was arranged alphabetically. The horizontal relationships were retained by listing the remaining words in their generic order. For example:

23–0	Agronomy
32–9	Agronomy-Tobacco
32–9	Tobacco-Agronomy-Tobacco
25–9	Agronomy-Tobacco-Production
25–9	Tobacco-Agronomy-Production
25–9	Production-Agronomy-Tobacco-Production
34–X	Agronomy-Tobacco-Production-Topping and Suckering
34–X	Tobacco-Agronomy-Production-Topping and Suckering
34–X	Production-Agronomy-Tobacco-Topping and Suckering
34–X	Topping and Suckering-Agronomy-Tobacco-Production-Topping and Suckering

Repetition of the first word as the last word in the horizontal line was necessary at times to keep the code numbers correct.

* Presented before the Division of Chemical Literature, ACS National Meeting, Washington, D. C., March 23, 1962

It was recognized that false retrievals would occur under the scanning system, since keywords were tabulated without any indication of their function or their relationship to other keywords. It was thought, however, that the problem of false retrievals could be minimized by detailing the question asked of the system. This proved to be wrong. A request for all reports concerned with interactions of items A, B, and C, for example, might bring a retrieval of many references to A, B, and C, but there was no way of determining, from the IBM cards alone, whether A, B, and C were concerned with each other or whether A was concerned with D, B with E, or C with F. To find the documents which were pertinent to the request required that the original reports be examined. We had neither the time nor the inclination to do this.

So the scanning system for encoding internal reports at the Philip Morris Research Center was discarded, and the development of a more serviceable system was begun.

The literature was examined to see how others had handled the problem of false retrievals, with attention centered particularly on articles by J. C. Costello, Jr.,[1] Don R. Swanson,[2] and Fred R. Whaley.[3]

It was decided that a variation of the collating system would be best for our purposes. In this system an IBM card is assigned to each keyword rather than to each document. At first the keywords were assigned code numbers. It was soon realized, however, that this simply added two steps to the system (one during encoding and one during retrieval) and that it added nothing to the effectiveness of the system. The use of code numbers was abandoned, therefore, and the use of the words themselves was adopted.

The decision to list role indicators, codes to show functional relationships between keyowords, prompted the inclusion of secondary keywords, which act on or are acted upon by the primary keywords. Swanson's hypothesis that "phrases two words in length" might be useful for "practical purposes of text searching"[2] suggested the inclusion of modifiers of the keywords.

Not every primary and secondary agent appearing in a document has a modifier but where such a modifier exists, its inclusion on the IBM card is often very helpful, and at times indispensable, in decreasing the number of documents which have to be considered in answering a retrieval question. For example, about 50% of cigarettes sold today are of the filter type. A great deal of research is underway on this part of the cigarette, and many internal reports concern themselves with different phases of this research. Therefore, it is most important from the retrieval standpoint that the appropriate modifier be used when "filter" appears as a keyword. Otherwise, false drops might result during a search concerning filters.

In our new system four fields, each consisting of thirteen columns, are laid out on the IBM card and assigned as follows: field I, columns 26–38, modifier of the secondary agent—that is, keyword; field II, columns 40–52, secondary agent; field III, columns 54–66, modifier of the primary agent; and field IV, columns 68–80, primary agent (Fig. 2).

A word which is longer than thirteen letters and, therefore, too long for its assigned field on the IBM card, is abbreviated, and the abbreviation is listed beside the word in the descriptor dictionary. This abbreviation is the "raison d'être" of the dictionary, for without the need for a source of standardized abbreviations there would be no real need for a dictionary. Synonyms have proven to be of minor concern.

The following illustrates how the dictionary is set up:

Anise, Oil of	Use Carvone
Aromatic hydrocarbon	Use Hydrocarbon, aromatic
Aluminum chloride	Alum. Chloride
Carbon dioxide, Evolution of	CO_2, Evolut. of
Combustion tube	Use Tube, Combustion
Hydrocarbon, Aromatic	Hyd. car., Arom
Silver ammonium chromate	Sil. NH_4 Chro.
Tube, Combustion	Tube, Combust.

Descriptors which do not need to be abbreviated are kept in a different listing.

It is necessary to use the reverse form of a keyword combination, that is, to put the generic term first when the keyword term consists of more than one word, only when that keyword falls in the primary agent field of the document card. This is because the cards are filed alphabetically by the primary agent.

Columns 21–22 of the IBM card are reserved for role indicators, designated by numerical code, which show the relationships between descriptors. Six roles have been found to be sufficient for the Research Center internal reports. They are defined as shown in Table I.

Table I

(A refers to the primary agent, B to the secondary agent.)

01 Relationship between (A and B), comparison of (A with B)
02 Preparation, formation, generation of (A from B or A by means of B), development of (A from B), design of (A)
03 Analysis for, measurement of, presence of, evaluation of (A in B)
04 Chemical or physical treatment of (A with or by B)
05 Formulation for, composition of (A)
06 Analysis of (A for B)

Roles 03 and 06 are opposites. The listing of 06, when 03 is applicable, is necessary because the IBM cards are filed alphabetically by primary agent, that is, by A. To find B would require that all the cards in the file be sorted, if the relation between A and B were not reversed so that B, in turn, becomes A and is shown as a primary agent. Similarly, under role 01, A and B are of equal weight. It is necessary, therefore, when role 01 applies, to punch two cards, showing each term as the primary agent. This is not necessary, of course, when the primary and secondary agents are the same as, for example, in a comparison (01) between filtered cigarettes and unfiltered cigarettes.

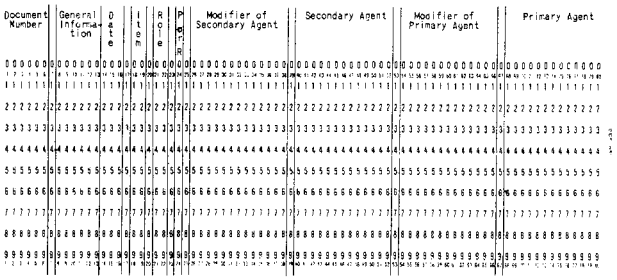
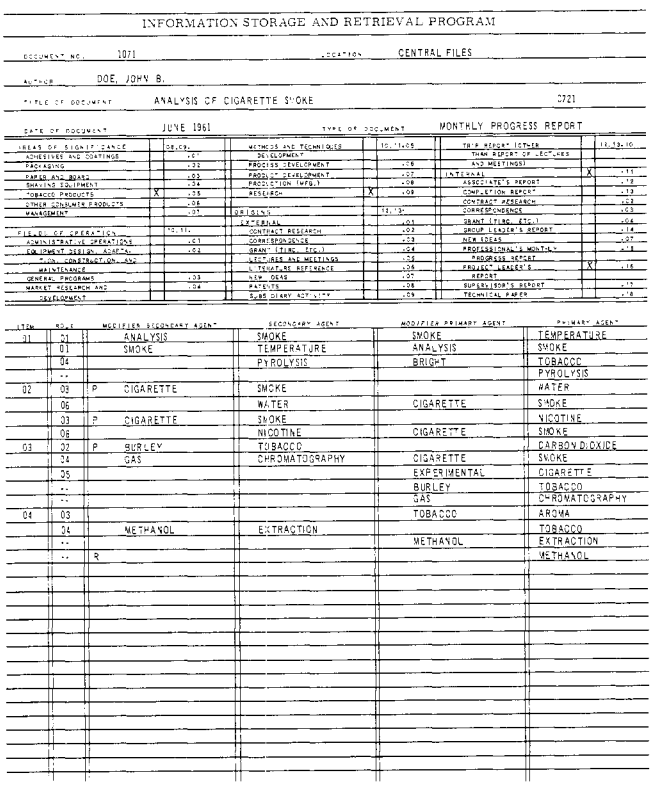


Fig. 2.

Numerical punches in columns 18–19 indicate the portion, or item, of the document which is being encoded. Reports are divided into items according to the subject matter being discussed. A document might consist of one item or a half-dozen or more items, depending on how many different subjects are touched upon.

Column 24 is reserved for a letter, P or R, where applicable, which denotes the function of a primary agent which is a named compound. P indicates that the compound is a product which is present naturally or as the result of a chemical reaction. R denotes that the compound is a reagent, an additive of some kind.

Not all agents have role indicators assigned to them since it is sometimes difficult to establish their relation to other agents. Such agents are included in the encoding, however, because they do qualify as keywords and could serve to locate documents pertinent to an information request.



INFORMATION STORAGE AND RETRIEVAL PROGRAM

DOCUMENT NO. 1071 LOCATION CENTRAL FILES

AUTHOR DOE, JOHN B.

TITLE OF DOCUMENT ANALYSIS OF CIGARETTE SMOKE 0721

DATE OF DOCUMENT JUNE 1961 TYPE OF DOCUMENT MONTHLY PROGRESS REPORT

AREAS OF SIGNIFICANCE		OBJECT	METHODS AND TECHNIQUES		THIS REPORT (OTHER THAN REPORT OF LECTURE AND MEETINGS)		
ADHESIVES AND COATINGS		.01	DEVELOPMENT		INTERNAL	X	.11
PACKAGING		.02	PROCESS DEVELOPMENT		ASSOCIATE'S REPORT		.12
PAPER AND BOARD		.03	PRODUCT DEVELOPMENT		COMPILATION REPORT		.13
SHAVING EQUIPMENT		.04	PRODUCTION (MFG.)		CONTRACT RESEARCH		.03
TOBACCO PRODUCTS	X	.05	RESEARCH	X	CORRESPONDENCE		.03
OTHER CONSUMER PRODUCTS		.06	ORIGIN		GRANT LETTING, ETC.)		.04
MANAGEMENT		.07	INTERNAL		GROUP LEADER'S REPORT		.14
FIELD OF OPERATION			CONTRACT RESEARCH		NEW IDEAS		.07
ADMINISTRATIVE OPERATIONS		.01	CORRESPONDENCE		PROFESSIONAL'S MONTHLY PROGRESS REPORT		.08
EQUIPMENT DESIGN, ADAPTATION, CONSTRUCTION, AND MAINTENANCE		.02	GRANT LETTING, ETC.)		PROJECT LEADER'S REPORT	X	.16
GENERAL PROGRAMS		.03	LECTURE AND MEETINGS		SUPERVISOR'S REPORT		.17
MARKET RESEARCH AND DEVELOPMENT		.04	LITERATURE REFERENCE		TECHNICAL PAPER		.18
			NEW IDEAS				
			PATENTS				
			SUBSIDIARY ACTIVITY				

ITEM	ROLE	MODIFIER SECONDARY AGENT	SECONDARY AGENT	MODIFIER PRIMARY AGENT	PRIMARY AGENT
01	01	ANALYSIS	SMOKE	SMOKE	TEMPERATURE
	01	SMOKE	TEMPERATURE	ANALYSIS	SMOKE
	04		PYROLYSIS	BRIGHT	TOBACCO
	--				PYROLYSIS
02	09	P CIGARETTE	SMOKE		WATER
	06		WATER	CIGARETTE	SMOKE
	03	P CIGARETTE	SMOKE		NICOTINE
	08		NICOTINE	CIGARETTE	SMOKE
03	02	P BURLEY	TOBACCO		CARBON DIOXIDE
	34	GAS	CHROMATOGRAPHY	CIGARETTE	SMOKE
	05			EXPERIMENTAL	CIGARETTE
	--			BURLEY	TOBACCO
	--			GAS	CHROMATOGRAPHY
04	03			TOBACCO	AROMA
	04	METHANOL	EXTRACTION		TOBACCO
	--			METHANOL	EXTRACTION
	--	R			METHANOL

Fig. 3.

When a specific chemical compound is named as a primary agent, it is entered not only in the document card file but also in a file devoted to the definitions of chemical compounds. In this latter file each compound is defined as to its class, whether it is an aldehyde, a ketone, a hydrocarbon, an acid, *etc.*, as to whether it is aliphatic, aromatic, alicyclic, or heterocyclic, organic or inorganic, saturated or unsaturated, and as to whether it contains elements other than carbon, hydrogen, and oxygen. This card file makes possible, by machine sorting, quick determination of all compounds in any of these categories. The value of this may be seen from an example:

A request might be made for a listing of all the aromatic aldehydes which have been found by our laboratory to be present in cigarette smoke. Sorting of the compound definition cards for aldehyde and aromatic drops out the cards of those compounds which fall into these categories. Reference to the document cards shows in which reports these compounds have been related to smoke. The letter P or R in column 24 indicates whether the compounds are products found in smoke or reagents added to smoke.

Columns 1–6 of the document card are reserved for the assigned report number; columns 8–16 are used for coded general information concerning the report. The Area of Significance discussed in the report is punched into columns 8–9. This indicates whether the document is concerned with tobacco products, shaving equipment (American Safety Razor Company is a subsidiary of Philip Morris Incorporated), packaging (Milprint, Inc.), industrial adhesives (Polymer Industries, Inc.), paper (Nicolet Paper Corporation), or matters which would properly come under the term "management." In columns 10–11 punches indicate what Field of Operation is involved, research, process development, equipment design, or any of a half-dozen other possibilities. The origin of the report is shown in columns 12–13; and the date of the report, month and year, is shown in columns 14–16.

A typical worksheet as it might be filled out during the encoding of a Research Center report is shown in Fig. 3.

Seventeen IBM cards are needed to record the information on this work sheet, one for each of the primary agents and its accompanying terms and codes. The document number and date and the checked items in the top part of the sheet are manually punched into the first card, then automatically duplicated by programming in each of the succeeding cards. Fig. 4 illustrates how the cards look when punched.

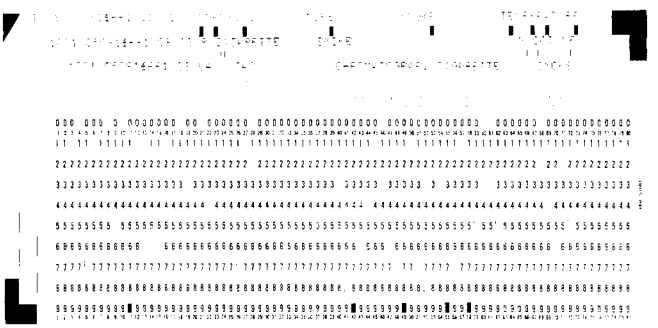


Fig. 4

After the cards are punched and verified, they are filed alphabetically according to the primary agents with cards from other encoded reports. New words and, where needed, their standardized abbreviations, are added to the vocabulary file.

It is difficult to estimate the average length of time required to encode a report by the described system. Reports vary greatly in the complexity of relationships within their subject matter. A document which discusses a single subject, and only one or two aspects of that subject, might be read and encoded in five to ten minutes. Another document, which discusses several subjects and several facets of those subjects, would require considerably longer to read and encode.

The reward for the deep-indexing and careful notation of relationships comes in the retrieval part of the system. When a request for information is received, the question

is read carefully and the keywords and their relationships are determined. The document cards are then consulted directly, without need for reference to a code dictionary. Quite often the report or reports which will answer the question can be found by reference to just one of the keywords in the document card file because the other keywords involved are present as secondary agents and modifiers. The relationships of these latter descriptor terms to the primary agent are immediately apparent from the coded role indicators.

As was mentioned, there are instances in which keywords are listed without role indicators being assigned. This might occur in the naming of a piece of apparatus used, a named method employed, or a compound tested. It occurs when role indicators 02 and 04 are employed. The secondary agents are relisted as primary agents. When such a keyword appears in a question, it is necessary only to match document and item numbers to relate it to other keywords. This can be done by visual inspection or, if the number of IBM cards is large, by machine sorting. Probably the quickest method is to sort the cards by machine into numerical order by document numbers and then to complete the matching by visual inspection.

The numbers of the reports which appear, from the information on the IBM cards, to be pertinent to the request are recorded along with the dates of the reports and the appropriate item numbers. The dates are needed because the reports are filed in binders by months, and the indexing has not been done in chronological order. The pertinent reports are taken from their various binders, the pages containing the appropriate items are machine duplicated, and copies are sent to the person who requested the information.

When the information retrieval program was first set up for the Philip Morris Research Center using the scanning system of indexing, the documents being encoded included articles in journals as well as laboratory reports. At that time an index card file was begun which was arranged in numerical order and showed the physical locations of the documents. When the encoding method was changed to the collating system of indexing, it was decided to concentrate on the internal reports and to leave until later the information from outside sources. This decision made the index card file superfluous since the internal reports are kept in a central location.

It is hoped that in the near future filing of the reports can be changed from binder filing by months to loose filing in numerical order according to the assigned document numbers. This would make it a simple matter to pull out reports which are considered pertinent to a request.

The greatest disadvantage of the system described lies in the number of IBM cards involved. For a large file system this would probably be prohibitive because of the storage problem. For a small system, however—fewer than 10,000 internal reports have been written since the Philip Morris, Incorporated, Research Department was organized in 1952—the advantages far outweigh this disadvantage.

The advantages can be summarized as follows: (1) Because of the large amount of information contained on each document card, false retrievals are virtually eliminated. (2) Since the keywords themselves are punched into the document cards, there is no need for a *coded* descriptor dictionary. (3) The number of role indicators used is small, being six, and the code is easily learned. (4) The amount of machine sorting required is small. The 082 Sorter is used only when there are a large number of cards pertaining to a given primary agent and when machine sorting would be faster than manual sorting for a desired report number, date, role indicator, secondary agent, or modifier. (5) The use of modifiers enhances the information conveyed by the agent words. (6) Each document card, because it contains the report number, date, and item term, pinpoints the source of the information which it contains.

REFERENCES

(1) Costello, J. C., Jr., "Uniterm Indexing Principles, Problems and Solutions," *Am. Document.* **12**, 20–26 (1961).

(2) Swanson, Don R., "Research Procedures for Automatic Indexing," paper presented at the American University Third Institute on Information Storage and Retrieval, Washington, D. C., Feb., 1961.

(3) Whaley, Fred R., "A Deep Index for Internal Technical Reports," J. H. Shera, A. Kent, and J. W. Perry (ed.) "Information Systems in Documentation," Interscience Publishers, New York, N. Y., 1957.